

Building State Capacity in Stages: Brokers in Public Service Delivery

Torsha Chakravorty[†]

Parminder Pal Singh Sandhu[‡]

WIP – Please do not cite or circulate.

Abstract

Citizens need to establish eligibility for public benefits, and many pay private agents (brokers) to help obtain them. We study how this broker market responds when the government makes it easier to apply directly. A staffing reform in a large Indian state added workers who help submit applications, while leaving approving officials unchanged. Using pre-existing differences in district exposure to the reform, we find that a 1 percent staffing increase raises broker-mediated applications by 1.7 percent, compared to 1.1 percent for direct applications, and induces broker entry. Yet direct applications become faster, reducing brokers' average speed advantage. These patterns rule out a model where public workers purely substitute for brokers and, under maintained assumptions, one where brokers only obtain priority from approvers. A model where public workers make brokers more productive can explain joint results. Our findings highlight the role of informal markets in building state capacity.

Keywords: State capacity; bureaucratic intermediaries; public service delivery; administrative burdens; informal markets; India.

JEL codes: D73, H11, O17.

*Corresponding author: tchakravorty@ucsd.edu

[†]I am grateful to my advisors, Paul Niehaus, Julie Cullen, Karthik Muralidharan, Prashant Bharadwaj, and Sam Bazzi, for their guidance, advice, and many helpful comments. I thank participants at the RIDGE May Political Economy Conference, the BEDI Conference at the University of Pittsburgh, and the Development Advising Group at UC San Diego for valuable feedback. I especially thank Aaditya Dar, Aprajit Mahajan, Prabhat Barnwal, Emily Aiken, Meera Mahadevan, Colin Sullivan, Yasir Khan, and Sayantan Roy for their many thoughtful comments and conversations about this project over the past three years. I am indebted to the many government officials who took time from demanding schedules to speak with me about their work and this project, and who were extraordinarily helpful during my fieldwork. All errors are my own.

[‡]Government of India. The views expressed here are those of the authors and do not necessarily reflect those of the Government of India.

1. Introduction

Formal entitlements to welfare programs often do not guarantee de facto access. Citizens must document their eligibility in order to access public benefits. This can require learning unfamiliar procedures, assembling supporting documents, making repeated visits to government offices, and waiting for officials to make a decision. Therefore citizens often use private agents (brokers) to access public benefits and welfare schemes. They are unregulated private intermediaries who charge for help to navigate government procedures. Brokers can save time, reduce application mistakes, and help citizens complete applications that might otherwise take a long time. However, they may also help clients bypass procedures or obtain preferential treatment from officials. The distinction matters for how we understand their role in public service delivery. Assistance can reduce the real costs of accessing the state, while preferential treatment can benefit one applicant at the expense of others when official attention is scarce. These certificates are often required to access public benefits. Understanding access to the state therefore requires understanding both how citizens obtain these certificates and the private markets that emerge around that process.

A common policy view is: brokers exist and are widely used because governments make public services difficult to access directly¹. Citizens hire brokers when they cannot easily get help from the state, so governments often try to reduce reliance on them by making direct access easier. More public assistance can replace the help brokers provide in preparing applications, answering questions, and correcting mistakes. Faster and more reliable service can also reduce the value of paying someone to navigate delays. Under this view, expanding public capacity should reduce citizens' demand for brokers. As direct access improves, fewer citizens should need an intermediary. Public capacity and private intermediation are therefore substitutes. Expanding the state should crowd out at least part of the private market that has formed around it.

We test this belief using a statewide staffing reform in a large Indian state. Citizens in our setting apply for government-issued certificates that establish income, caste, residence, birth, death, and other forms of eligibility. These documents are often required to access welfare programs, public benefits, and government jobs. Citizens submit their applications through a statewide network of public service centers known as *Sewa Kendras*. In March 2023, the government sharply increased staffing at these centers. The reform expanded the state's capacity to help citizens prepare and submit applications, but it did not increase the capacity of the officials who reviewed and approved them. It therefore made the application stage substantially easier without expanding

¹This view is explicit in many administrative reforms. For instance, India's Direct Benefit Transfer system was designed to send welfare benefits directly to recipients and reduce intermediaries and leakage. Similarly, government guidance for Jan Seva Kendras describes application procedures as being designed so citizens can complete them without the help of "touts or middlemen."

the downstream capacity that ultimately determines how quickly a certificate is issued. An outline of the application system is in Figure 1.

Brokers can help citizens at both stages of the process. At the application stage, Sewa Kendra operators help citizens assemble documents and submit applications, while brokers can provide many of the same services privately. At the approval stage, government officials based in separate district or block-level offices review applications, verify eligibility, and authorize certificates. Brokers may also create value at this stage by helping clients obtain faster or preferential treatment once an application enters the system and reaches an approver. The staffing reform therefore changed the value of different broker services in different ways. More operators could make brokers less valuable for preparing and submitting applications. But since the reform did not expand approval capacity, it affected the net value of brokerage ambiguously. The effect of greater state capacity on the broker market therefore depends on where in the production process brokers create value.

The administrative records let us observe both how the state delivers services and the private brokerage market that forms around it. For every application, we observe submission and issuance dates, service type, location, and the mobile number attached to the application. Most mobile numbers appear only once, but a smaller set recur across applications submitted for different citizens. We use this repeated activity to identify likely intermediaries. In our baseline measure, we classify a mobile number as a broker if it appears on at least three applications during our entire November 2022-August 2023 sample. Because citizens typically need these certificates only occasionally, frequent reuse of the same number is unusual and is also viewed by government IT officials as a marker of intermediary activity. This measure will miss brokers who submit applications using their clients' own numbers and may misclassify shared family numbers or unpaid helpers as brokers. We interpret it as a proxy instead of a complete census of brokerage. If these classification errors are unrelated to application outcomes, they will generally attenuate measured differences between brokered and direct applications.

The statewide hiring expansion generated larger increases in application-stage capacity in some districts than in others for predetermined reasons. We exploit this variation by interacting pre-reform staffing per capita with the subsequent statewide increase in hiring. Districts with greater pre-reform staffing were more exposed to the common expansion because the government allocated new operators through the existing service-center infrastructure. We use this exposure measure to instrument for actual staffing per capita and compare changes in applications, broker activity, and processing outcomes across more- and less-exposed districts. The design controls for

district fixed effects and common seasonal patterns. Identification requires that, absent the reform, districts with different exposure would not have experienced systematically different changes in these outcomes. We test this assumption by asking whether exposure predicts differential trends before the staffing expansion and by examining other policy changes that could have affected application demand or brokerage during the same period.

Contrary to the simple crowd-out prediction, additional application-stage staffing expands the broker market disproportionately. A 1 percent increase in staffing raises broker-mediated applications by approximately 1.7 percent, compared to 1.09 percent for applications filed directly by citizens. This is not simply greater activity among existing brokers. The number of active likely brokers also rises with staffing, consistent with entry into the market. At the same time, direct applications increase. Thus, the reform expands access through both channels rather than merely shifting applications from one to the other. The larger response through brokers raises their share of application activity. Greater public capacity therefore crowds in both the use and the supply of private intermediation.

Brokerage expands even as the reform erodes a key visible advantage over direct access. Before the reform, brokered applications reached issuance about 4 days faster on average, than applications filed directly by citizens. Additional staffing substantially narrows this gap, primarily because direct applications become faster. Yet broker use continues to rise as this observed processing-time advantage shrinks. This pattern is difficult to explain if citizens hire brokers mainly to avoid a slow direct channel. One possibility is that additional public workers make brokers more productive by reducing the time and effort required to prepare and submit applications. If so, brokerage can expand even as direct service improves and the recorded processing-time advantage of using a broker falls.

The joint responses of broker use, entry, and processing times point to complementarity between public workers and brokers. We formalize this intuition using three benchmark models. The first captures the original substitution view we introduced earlier: if brokers mainly help citizens overcome weak public assistance, expanding state capacity should reduce their value and lower the brokered share of applications. That prediction contradicts disproportionate expansion of brokered applications in the data. A second model allows brokers to obtain priority from approving officials but leaves their application-stage productivity unchanged. Under our maintained assumptions, priority alone cannot reproduce the joint increase in brokered applications and improvement in direct processing times. However, a simple complementarity model does rationalize this: additional application-stage workers reduce the time and effort brokers need to

serve each client, allowing broker entry and brokered applications to rise even as direct service improves. This does not imply that brokers never receive preferential treatment; rather, the response to the staffing expansion requires a productive complementarity between public capacity and private intermediation that preferential treatment alone cannot provide.

Productive assistance at the application stage can coexist with a broker advantage later in the process. We find little evidence that *Sewa Kendra* operators themselves give brokers priority: brokered applications are no more likely than direct applications to enter the system earlier in the day. This is the main margin through which these workers could move brokers ahead. The pattern changes once applications reach approving officials. We compare approval before the statutory deadline, the service-specific legal time limit for issuing a certificate, across services that give these officials more or less discretion. After the staffing expansion, the broker advantage in timely approval shrinks substantially for lower-discretion services but remains essentially unchanged for higher-discretion services. This persistence is consistent with brokers retaining an approval-stage advantage precisely where officials have greater scope to influence outcomes. The evidence does not establish how that advantage is obtained, but it leaves room for preferential treatment, including the possibility of payments for faster approval.

Administrative burdens are shaped not only by government procedures, but also by private markets that emerge to help citizens navigate them. A large literature shows that information, paperwork, and application assistance can substantially affect take-up of public benefits (Bettinger et al., 2012; Bhargava and Manoli, 2015; Finkelstein and Notowidigdo, 2019; Banerjee et al., 2021; Gupta, 2026). Much of this work examines interventions in which governments, researchers, or other organizations reduce these frictions directly, rather than how assistance is supplied through private markets. Yet private participation can itself respond to public policy: higher SNAP spending during the Great Recession, for example, led more retailers, particularly non-grocery chains, to join the program (Byrne et al., 2025). We study a different form of private participation: intermediaries whom citizens pay to obtain government documents needed to access social welfare schemes and other services. Expanding frontline government staffing makes it easier for citizens to apply directly, but it can also make intermediaries more effective at assisting applicants. We find that both direct and intermediary-mediated applications increase, alongside entry by intermediaries. The private market for assistance can therefore expand even as the government makes the application process easier. This response suggests that the costs of accessing the state depend not only on government procedures, but also on how the private market for navigating them adjusts.

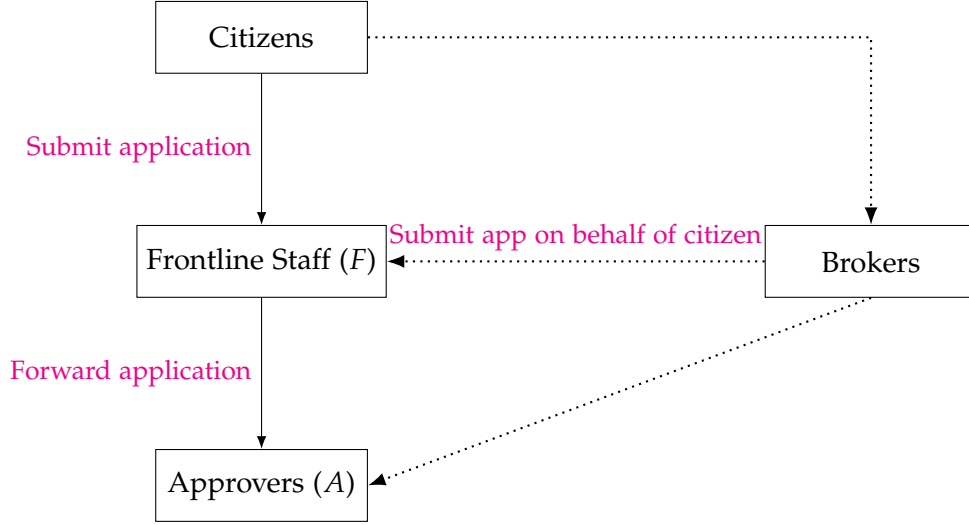
Bureaucratic intermediaries can complement public workers even while retaining advantages where official discretion matters. Existing work often emphasizes intermediaries who facilitate corruption or help citizens circumvent official procedures ([Bertrand et al., 2007](#); [Chalendard et al., 2023](#)), while [Fredriksson \(2014\)](#) shows that intermediaries can also save citizens time in settings with corruption and red tape. Policies aimed at intermediaries can also reduce access to public programs ([Meckel, 2020](#)). This paper identifies another possibility: productive complementarities between public workers and intermediaries. When the state expands application-stage capacity, broker use and entry increase rather than decline, consistent with public workers making brokers more productive. At the same time, brokers retain an advantage in timely approval in services where officials exercise greater discretion. Productive assistance and preferential access can coexist within the same intermediary market. Expanding public capacity need not reduce broker activity. When brokers complement public workers, greater public capacity can increase demand for their services, even as brokers retain advantages in discretionary approvals.

The effects of building state capacity depend on where governments add capacity and on how private actors respond. A large literature studies how government performance depends on the recruitment and quality of public employees, the organizations in which they work, and the resources available to them ([Dal Bó et al., 2013](#); [Muralidharan et al., 2023](#); [Ashraf et al., 2020](#); [Romero and Singh, 2022](#); [Best et al., 2023](#); [Ganimian et al., 2024](#); [Aneja and Xu, 2024](#); [Barnwal et al., 2026](#)). Related work shows that actors outside the formal bureaucracy can also shape state performance: [Balán et al. \(2022\)](#), for example, show that local chiefs improve tax collection by supplying information and local knowledge that state agents lack, even while collecting more bribes. We emphasize a different margin: government services are often produced in stages, and workers at different stages need not be interchangeable. In our setting, the state expands capacity for preparing and submitting applications while leaving approval capacity essentially unchanged. This improves direct service, but it also changes the costs and incentives of brokers who operate around the same administrative process. Additional public workers at one stage can therefore expand a private market rather than displace it, while bottlenecks at another stage continue to shape the value of intermediation. Building state capacity thus requires understanding both how government production is organized across stages and how private markets adjust around it.

The remainder of this paper is organized as follows. Section 2 outlines the institutional background and the staffing reform, section 3 introduces the data and variables of interest, section 4 details our empirical strategy, and section 5 presents descriptive evidence on how our outcomes of interest evolved following the reform. Section 6 reports the IV results and estimated elasticities, with Section 7 presenting the benchmark models. Finally, Section 8 discusses implications for

citizen welfare, and Section 9 concludes.

Figure 1. Two-stage production process



2. Institutional Background and Staffing Reform

In this section we describe how an application becomes a certificate, role of different staff at the two production stages (application submission and approval), and the staffing reform underlying our empirical strategy of March 2023. We use confidential administrative data from a large Indian state with a population of roughly 32 million. We consider this state to be an ideal setting to examine our question because the state has invested in a unique digital governance infrastructure called the *Sewa Kendras* (SKs).

2.1. Sewa Kendras and the application process

An overview of SKs. These are service centers acting as an interface between the citizen and the state. They simplify the process of filing applications by acting as centralized kiosks for submitting applications, general queries, grievances, and miscellaneous post-office related work pertaining to all departments. Citizens can use these kiosks to access basic services like certificate issuance, record updating, and license and permit applications. There are 516 SKs in the state, each catering to roughly 45,000 citizens. First inaugurated in August 2016, there were more than 2,000 SKs before 2017 with very low footfall (providing only 170 services). After 2017, existing *Suvidha* Centers, *Saanjh Kendras*, and others were amalgamated because they were not financially remunerative, reducing the number to 516 SKs.

Services provided. Currently SKs deliver over 300 government services, of which about 30 percent are ‘online services’, i.e. entire process of application, examination, approval and delivery is managed electronically (and in fact no physical records or files are maintained). While citizens in theory can complete an application online, in practice nearly all citizens default to submitting their applications in-person at the SK because of recurring technical issues on the online portal. The certificates we study in this paper are a subset of the services that the SKs provide. Our analysis focuses on the subset of services for which we received complete application-level administrative records, including submission and issuance dates. These include income, caste, residence, birth and death certificates, pensions, and related administrative approvals across six government departments. Besides identity certificates, services like updating demographic and biometric information, land mutation, driving license updates, among others are also processed by SK operators.

Types of SKs. There are three types of SKs in the state: type-I SKs are located in district headquarters with 12 counters on average; type-II SKs cater to urban population with 4 counters; and type-III SKs cater to rural population with 2 counters.

How SKs work. As soon as they arrive, citizens can get a token number through the helpdesk at the SK. They wait for their turn in the designated waiting area until their token number is displayed/announced by a SK operator. Citizens then proceed to the respective counter that called their token number and interact with the SK Operator to submit their application. After successfully submitting the application, the operator provides the citizen with a fee receipt. The official fee for many services are near zero but a few services have a maximum facilitation fee of about INR 50-70 (equivalent to USD 0.52-0.73). Using the unique application ID generated, citizens can track the status of their applications online. Operators usually forward the application to approving officials who verify eligibility and all documents and approve applications. Once the application is reviewed and approved, citizens can collect the final certificate by visiting the same SK they originally submitted the application.

SK operators. All SK operators (more than 50 percent female) hold at least an undergraduate degree. Many type-I and II operators were private sector employees before they joined, unlike operators in type-III SKs. They receive recurring software and computer training for professional development. Citizens rely on these frontline operators to assemble, fill out, and informally verify documents before they are forwarded to department officials. The operators put together an applicant’s package, enters all application details in the online database and forwards the

application to a higher government official who is responsible for verifying and approving the application. Figures 2 and 3 illustrate the frontline operator set-up. In March 2023, the policy reform we study is a statewide increase in their hiring.



Figure 2. Citizen-facing SK counters



Figure 3. Behind SK counters with operators

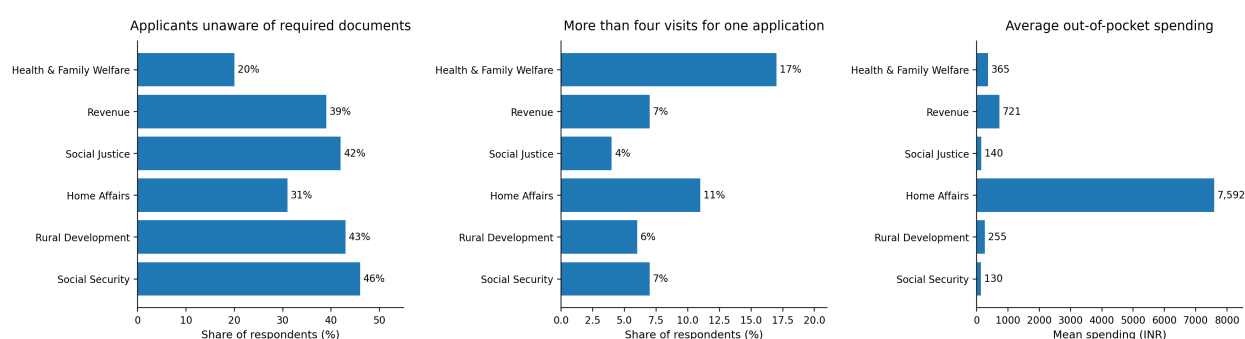
Approving officials. After a SK operator submits an application, it is sent to the government department responsible for that service. Officials in these departments verify the applicant's eligibility and supporting documents and decide whether to approve, or send the application back to be redone, or in very rare cases (less than 0.1 percent), reject. These officials are separate from SK operators and are generally based in district, sub-district, departmental, or other government offices rather than at the service centers themselves. The organization of approval varies across services. Some departments, such as revenue, health, and policing, have extensive local offices and staff, while others rely on a smaller number of district-level or centralized officials. Approving officials also typically perform other administrative duties, so the number of officials formally assigned to a department does not necessarily measure the capacity available for processing these applications. Vacancies are common in several departments, further limiting effective approval capacity. Each service is governed by the state's Right to Services Act, which specifies the official responsible for approval and the maximum time allowed for delivering the service. Applications may be rejected, but rejections are rare in our data. Importantly for our empirical design, the frontline staffing expansion increased the number of SK operators without a corresponding increase in these downstream approving officials.

2.2. Citizens' SK experience: visits and out-of-pocket expenses

Existing survey evidence points to substantial frictions in navigating these services. [Barnwal et al. \(2026\)](#) surveyed applicants across six government departments and find that 44 percent did not

know which documents were required before arriving at a Sewa Kendra, with department-level shares ranging from 20 to 46 percent. Some applicants also make repeated visits: 7 percent report visiting a Sewa Kendra more than four times for the same application, with the share ranging from 4 to 17 percent across departments. Out-of-pocket spending can also be substantial relative to the low official fees. Median reported expenditure is INR 180 overall, with department medians ranging from INR 40 to INR 5,000. These patterns suggest that obtaining a government certificate can involve information costs, repeated interactions with the bureaucracy, and meaningful private expenditure (Figure 4).

Figure 4. Survey results from Barnwal et al. (2026)



2.3. The staffing reform

On 10 March 2023, during the State budget, the government announced an allocation of roughly USD 7.4 million towards e-governance. They did not specifically outline how the money was to be spent. However, they did highlight their focus on improving public service delivery and filling up vacancies. We examine staff records from the state's administrative data and find no change to approver hiring rates in the sample period around the state budget. However, we do find a rapid growth in the number of operators in the system. Our conversations with officials managing the SKs assure us that this is because hiring operators is much cheaper and less politically charged than hiring higher government officials like approvers.

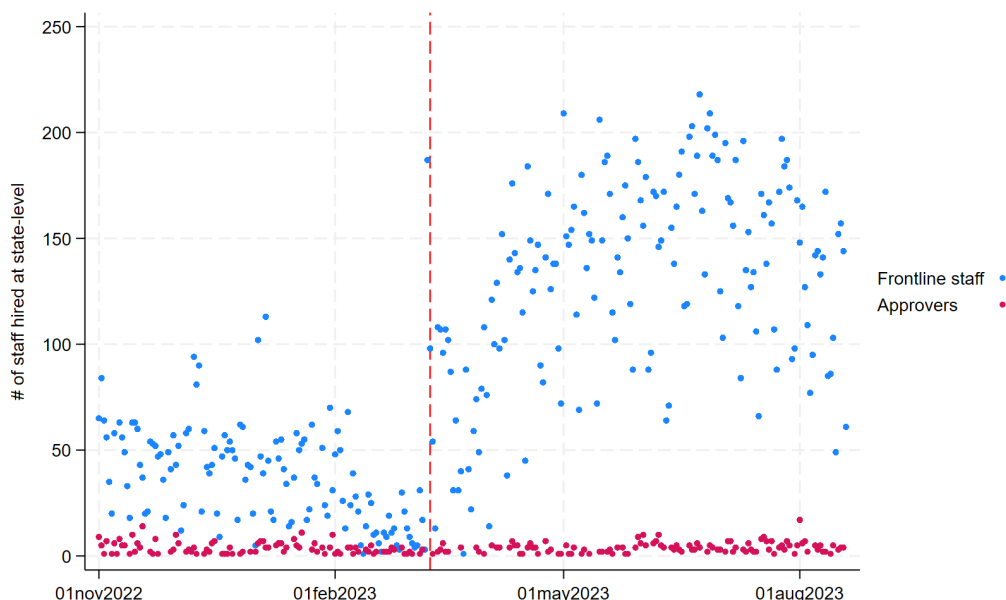
On March 10, 2023, the state government implemented a large staffing expansion across administrative offices (Figure 5). Staff onboarded on the government portal roughly tripled, from about 50 employees per day before the shock to about 150 employees per day immediately after the shock, with daily new staffing reaching roughly 175–200 employees by August 2023.

Further, since operators work out of SKs, assignment of operators are correlated with the existing SK infrastructure in the state. No new SK was opened. This meant that districts with

greater SK infrastructure, and therefore higher existing staff per capita, were also mechanically exposed more to the mass operator hiring at the state-level.

The time-series shock comes from a statewide staffing expansion during the sample period. Conceptually, we view this as the expansion of intake capacity only instead of a proportional expansion of the whole production process.

Figure 5. Staffing expansion and first-stage variation



3. Data and Measurement

3.1. Administrative application records

The state began digitizing the universe of all applications and their processing times from 2018 in an effort to increase transparency and improve public service delivery. Granular administrative records on applications and issuances of certificates for this state are not routinely available to researchers. This was a result of sustained relationship-building and trust with the government and department officials. The resulting data in this paper cover November 2022 through mid-August 2023. Earlier records were unavailable because political developments² made it difficult for the department to continue sharing data. We end the analysis mid-August because statewide floods led to (i) large-scale reassignment of government staff for flood relief from that date, and (ii) staff responsible for maintaining the administrative database were also affected, creating delays in the

²Farmers protests where the state faced severe disruptions and large-scale unrest.

updating of application records.

The data cover the universe of public service applications submitted through government service centers between November 1, 2022 and August 19, 2023. The sample period ends right when large floods hit the state in August 2023 leading to mass reallocation of government staff for flood relief.

Administrative data overview. The administrative data contains approximately 1.2 million applications across 24 districts and 6 government departments. The services include income certificates, caste certificates, residence certificates, pensions, birth and death certificates, etc and related administrative approvals. The administrative data record application submission dates, certificate issuance dates, service categories, district identifiers, and the mobile number associated with each application. Processing time is measured as the number of days between submission and certificate issuance. Processing times use subsequent approval dates, including approvals occurring after the submission period shown on the horizontal axis. Applications submitted near the end of the plotted period therefore remain observable beyond that period. We also compare processing times to service-specific statutory deadlines to construct delay outcomes. The main analysis aggregates outcomes at the district-by-date level.

Policy variable of interest. The key policy variable of interest for us is frontline operator staffing per capita at the district-date level. Let F_{dt} denote frontline operators per 100,000 residents in district d on date t , using 2011 Census population for the denominator. The administrative staffing data distinguish frontline service operators, who are the first point of citizen contact at application centers, from approving officials with greater authority over verification or approval.

3.2. Identifying brokered applications

Citizens routinely use brokers to file applications but since they are unregulated, administrative data do not record whether an application was filed through a broker. We infer likely broker activity from the repeated use of a single mobile number across applications. Most citizens apply for a given certificate rarely: some certificates are needed once, and others must be renewed about every five years. A citizen applying on their own behalf should therefore appear on few applications during our ten-month sample. Figure 6 shows the distribution of applications per mobile number. About 90 percent of numbers appear on only one or two applications, and the distribution has a long right tail. We classify an application as broker-mediated if its mobile number appears on three or more applications during the sample. Government officials and the

department's IT staff shared with us that they use the same rule to identify likely broker activity.

Misclassification of the broker proxy. Our broker measure is a proxy and can misclassify applications in either directions. It misses brokers who enter their clients' own mobile numbers. It can also count members of a household who share one number, or unpaid helpers who file for relatives and neighbors as brokers. If these errors are unrelated to application outcomes, they attenuate measured differences between brokered and direct applications. Additionally, our broker classification uses the full sample. This means a number with few applications before the reform can cross the threshold only because of applications filed afterward. To address this, we show that our results are similar when brokers are classified using pre-reform applications only, and with different thresholds for filed applications (Table 9). Another concern we consider is whether frontline operators themselves might be entering their own mobile numbers when filing for some applicants. We find this to be unlikely because our classified broker mobile numbers are observed to submit applications across multiple counters and service centers which is infeasible for a frontline operator (Figure 16).

Figure 6. High frequency mobile numbers identify likely broker activity



Notes: Each bar reports the percentage of all applications associated with mobile numbers that appear on the indicated number of applications over the study period. The dashed vertical line marks the classification cutoff: applications associated with mobile numbers appearing on three or more applications are classified as brokered.

3.3. Outcomes and variable construction

This classification allows us to construct brokered application volume, broker share of total applications, active brokers, and the direct-broker observed processing-time gap in district d at time t . The main broker-market outcomes are:

$$\text{Broker share}_{dt} = \frac{\text{Brokered applications}_{dt}}{\text{Total applications}_{dt}},$$

$$\text{Active brokers}_{dt} = \text{Number of high-frequency phone numbers active in } (d, t)$$

The processing-time outcomes include mean processing time, the fraction of applications delayed beyond statutory deadlines, processing time for brokered applications, processing time for direct applications, and the direct-broker time gap. The time gap is defined as

$$\text{Gap}_{dt} = \bar{T}_{dt}^D - \bar{T}_{dt}^B \quad (1)$$

where $\bar{T}_{dt}^D$ and $\bar{T}_{dt}^B$ are mean processing times for direct and brokered applications.

Observed processing time gap between direct and brokered applications is the mean difference in days from application submission to a certificate being issued by each of the two channels. We aggregate this at the district x date level. Table 1 shows substantial cross-district variation in processing times in the pre-reform period. In most districts brokered applications are processed faster on average, but the magnitude varies considerably and in a few districts the difference is small.

4. Empirical Design

4.1. OLS and the endogeneity of staffing

A simple OLS specification would estimate:

$$Y_{dt} = \beta F_{dt} + \nu_d + \tau_{dow} + \varepsilon_{dt}, \quad (2)$$

where Y_{dt} is an outcome in district d on date t , F_{dt} is frontline operators per 100,000 residents, ν_d are district fixed effects, and τ_{dow} are day-of-week fixed effects. The coefficient β measures the association between staffing and the outcome within districts, after controlling for day-of-week seasonality. However, OLS may be biased because staffing could respond to local demand, administrative backlogs, or other unobserved district-level conditions.

Table 1. Pre-shock applications, broker share and processing times by district

District	Average daily applications per 100,000	Brokered applications (%)	Nonbroker days	Broker days	Nonbroker minus broker	<i>p</i> -value
District 1	9.92	14.91	25.10	15.18	9.91	< 0.001
District 2	10.43	16.25	11.99	11.26	0.74	0.165
District 3	11.02	18.79	17.57	12.08	5.49	< 0.001
District 4	8.67	16.74	12.11	12.89	-0.78	0.108
District 5	8.46	13.03	9.64	7.81	1.84	< 0.001
District 6	12.74	25.37	12.15	9.06	3.09	< 0.001
District 7	4.86	15.86	18.25	16.49	1.76	0.001
District 8	8.67	27.65	3.44	2.85	0.59	< 0.001
District 9	12.80	19.72	5.02	3.65	1.37	< 0.001
District 10	8.06	15.56	11.91	9.13	2.79	< 0.001
District 11	12.36	21.75	9.91	9.97	-0.06	0.843
District 12	5.61	15.94	14.63	15.55	-0.91	0.008
District 13	12.29	28.97	14.55	10.79	3.76	< 0.001
District 14	8.84	13.59	22.94	12.69	10.25	< 0.001
District 15	17.57	23.38	8.37	8.89	-0.52	0.014
District 16	8.59	15.82	22.12	18.55	3.57	< 0.001
District 17	13.01	24.14	8.00	5.57	2.44	< 0.001
District 18	9.65	17.02	13.82	12.37	1.45	0.005
District 19	14.60	20.32	18.78	12.16	6.63	< 0.001
District 20	10.89	22.77	16.77	11.32	5.45	< 0.001
District 21	11.36	18.80	17.04	9.17	7.87	< 0.001
District 22	14.06	35.28	19.77	7.21	12.56	< 0.001

Notes: We anonymize district names following out Data Use Agreement with the State Government. All columns describe applications created from Nov 2022 through 9 Mar 2023. Processing time is the time (in business days) taken from application filing to issuance. These are separated by direct and brokered applications. The difference is mean non-broker processing time minus mean broker processing time. Two-sided *p*-values test equality of the means using a separate application-level regression in each district, with an intercept and broker indicator and standard errors clustered by application creation date.

Realized frontline staffing may respond to local demand, administrative backlogs, or other conditions that also affect our outcomes. If we run a simple OLS regression, we may therefore confound effects of additional workers with unobserved conditions that led to their assignment. To address this concern, we instrument realized staffing with the interaction between pre-reform frontline staffing per capita and the statewide staffing expansion. This instrument uses predetermined differences in districts' exposure to the expansion rather than their realized staffing adjustments alone.

4.2. Instrument Variable

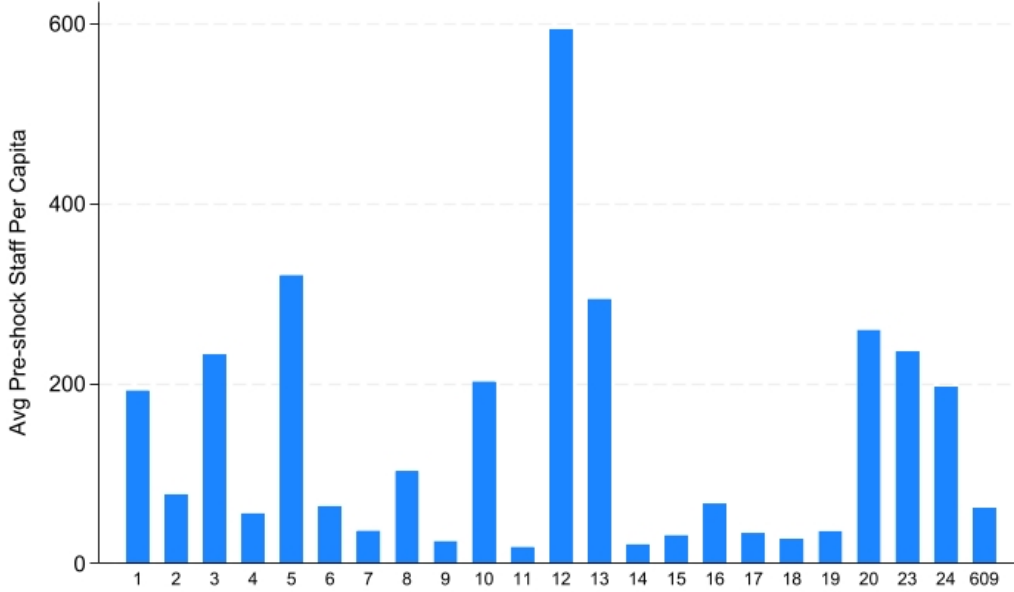
To address the OLS concerns, we use a differential exposure to a statewide staffing shock. The empirical design relies on differential district-level exposure (Figure 7) to a statewide increase in hiring of frontline operators (Figure 5) holding new hiring of approvers constant.

The instrument interacts a district's pre-shock staffing per 100,000 population with the aggregate statewide change in frontline staffing. Let $StaffPer100K_{d,pre}$ denote frontline staff per 100,000 residents in district d , averaged over Nov 2022 to 9 March 2023, using 2011 Census population. Let $\Delta AggregateFrontStaff_t$ denote the change in total frontline staff statewide from the mean pre-shock total frontline staff on day t . The instrument is:

$$Z_{dt} = StaffPer100K_{d,pre} \times \Delta AggregateFrontStaff_t. \quad (3)$$

$StaffPer100K_{d,pre}$ is fixed before the hiring expansion and does not respond to post-reform conditions in the district. $\Delta AggregateFrontStaff_t$ is common to all districts. If new staff are allocated in proportion to existing staff, Z_{dt} predicts the district's share of the statewide increase.

Figure 7. Staff per 100,000 population by district in 2021



The first stage is

$$F_{dt} = \gamma Z_{dt} + \nu_d + \tau_{dow} + u_{dt}, \quad (4)$$

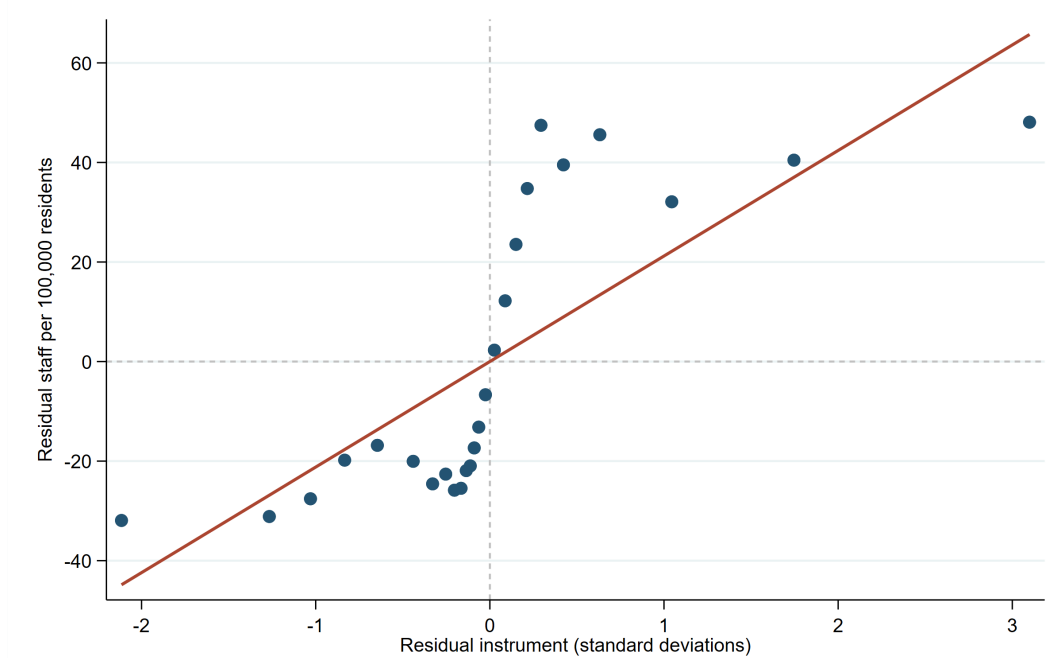
and the second stage is

$$Y_{dt} = \beta \widehat{F}_{dt} + \mu_d + \kappa_{dow} + \epsilon_{dt} \quad (5)$$

where $\widehat{F}_{dt}$ is the frontline staff per 100,000 residents in district d at time t as predicted by the instrument, ν_d and μ_d are district fixed effects, and τ_{dow} and κ_{dow} are day-of-week fixed effects. We cluster standard errors at the district and date level. The first stage has an F-stat of 15.9 and is visualized in Figure 8.

Figure 8 visualizes the first stage after accounting for persistent differences across districts and regular day-of-week patterns. The horizontal axis plots the instrument after partialling out district and day-of-week fixed effects. We divide this residualized instrument by its standard deviation: a one-unit movement along the horizontal axis corresponds to a one-standard-deviation increase in predicted exposure. The vertical axis is realized frontline staffing per 100,000 residents after removing the same district and day-of-week fixed effects. Thus, a value of 20 on the vertical axis means that realized staffing is 20 workers per 100,000 above the level predicted by the district and day-of-week effects. The positive relationship indicates that greater predicted exposure is associated with higher realized frontline staffing. A one-standard-deviation increase in the residualized instrument predicts approximately 21.2 additional frontline workers per 100,000 residents.

Figure 8. Instrument first stage



4.3. Identification assumptions and validity

The exclusion restriction. We need staffing expansion captured by the instrument affect applications, broker activity, and processing times only through its effect on frontline staffing in the district. For example, this restriction would fail if the expansion also brought changes in application requirements, service availability, or approval procedures that independently affected these outcomes. Instrument validity also requires that, conditional on district fixed effects and the included controls, predicted staffing be unrelated to other changes affecting the outcomes.

District fixed effects absorb persistent differences associated with pre-reform staffing, but do not absorb changing demand or administrative conditions associated with that exposure. Measuring exposure before the reform therefore does not, by itself, establish instrument validity.

It may still be the case that operators were being targeted at specific districts based on pre-shock performance. We verify this by plotting the relationship between district-level exposure and pre-shock outcomes like applications processed and processing times in Appendix A, Figure 14. We do not find any evidence of targeting based on pre-shock outcomes. We also verify that this shock to operator hiring is not a seasonal event that occurs every year.

Our preferred specification does not include date fixed effects. We use statewide timing of the expansion as well as differences in district exposure. Including date fixed effects substantially weakens the first stage by removing the common time-series component of the instrument. To retain the time-series variation, we need to assume other statewide developments do not generate changes in our outcomes that are correlated with the staffing expansion, conditional on the included controls. We assess these concerns using institutional evidence and placebo tests.

First, we check policy margins through which a concurrent change could most plausibly affect applications or processing times. Between November 2022 and August 2023, the government did not change the fees charged for services or for facilitation at SKs, the set of services offered, or the documentation required for any service. Second, we account for seasonal events like school admissions. Demand for caste and income certificates rises seasonally during the admissions period in December to February (before classes begin in April). However, this is in our pre-shock period and if anything, the effect of seasonality would imply that applications increase before the staffing shock happens, which is the opposite in our case under descriptive statistics below. Third, we test whether the timing of the reform is plausibly exogenous. We assign placebo reform dates within the pre-period and verify that they do not affect our service delivery measures (Appendix A, Figure 15).

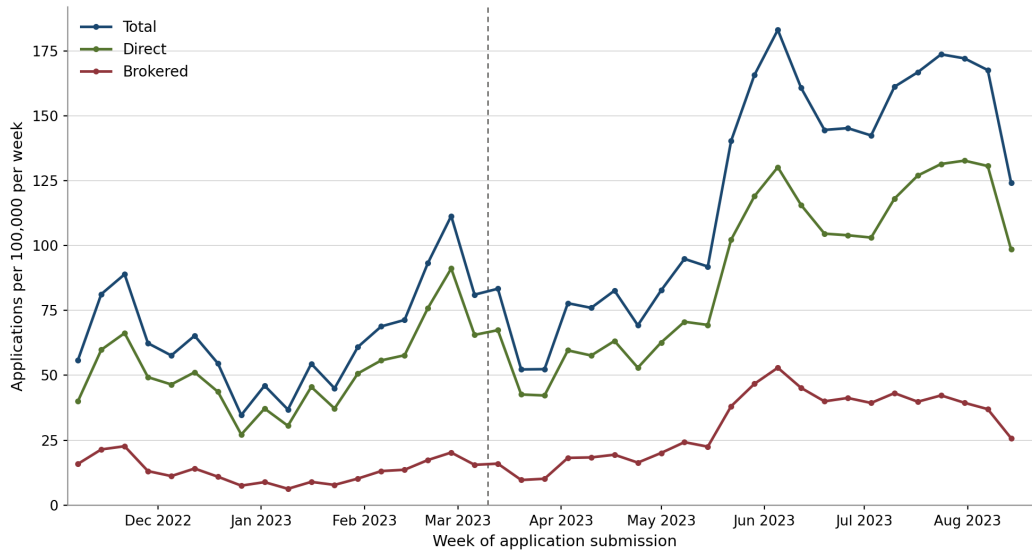
5. Descriptive facts around the staffing reform

5.1. Application volume and processing time

Before moving to the IV results, we show how our outcomes of interest evolved after the 10 March event. The raw time series graph plotting applications by their submission channel (direct or brokered) see a small dip around January for New Years typically being a slow time in the

bureaucracy, and in end of March because of a mobile internet suspension in the state for 2 days. However, the general expansion of applications per capita is clearly visible. We show evolution of total applications segregated by application route (direct and brokered). Applications per capita at the state-level rise from roughly 60 pre-shock to about 130 after.

Figure 9. Applications per 100,000 populations



5.2. Average processing time of applications

Mean processing time, i.e. difference in business days between submission and issuance, of applications improved after the event and is driven largely by direct applications. We plot this as a raw time series trend at the state-level in Figure 10. The time taken by brokered applications remain approximately similar throughout the sample period. This is important because the observed processing time advantage that citizens may have from applying through a broker appears to diminish after the frontline operator hiring shock. Intuitively, we would expect citizens to then reduce their demand for broker services if the only service brokers offered was guaranteed faster processing after application is submitted. But that is not the case, as we see both in the share of brokered applications above in Figure 9 as well as the number of active brokers per capita across the state, as is shown in Figure 11.

Figure 10 reports mean processing times by the week in which applications were submitted, separately for direct and brokered applications. The figure shows a gradual narrowing of the processing-time gap following the staffing expansion, largely reflecting a decline in processing times for direct applications.

Figure 10. Processing time by application channel

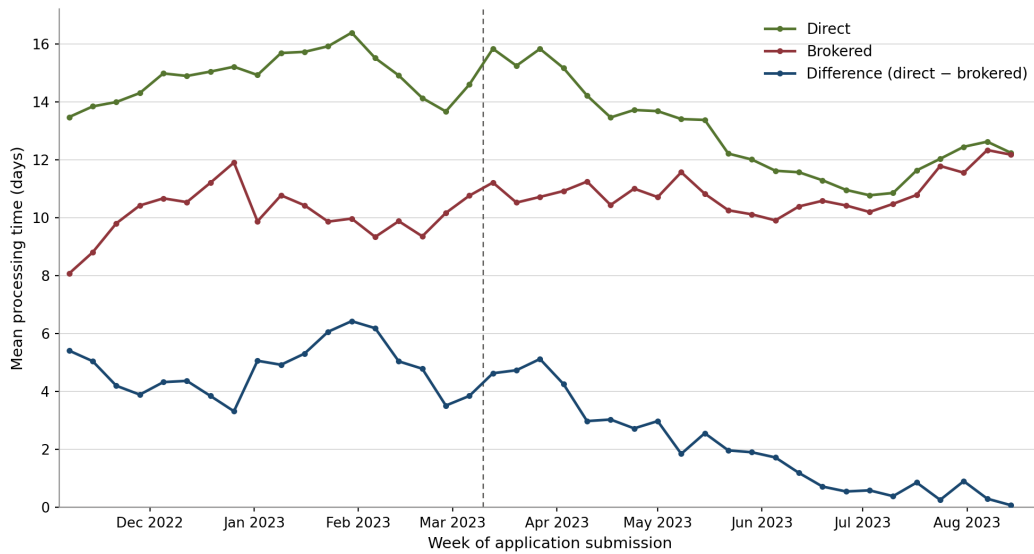
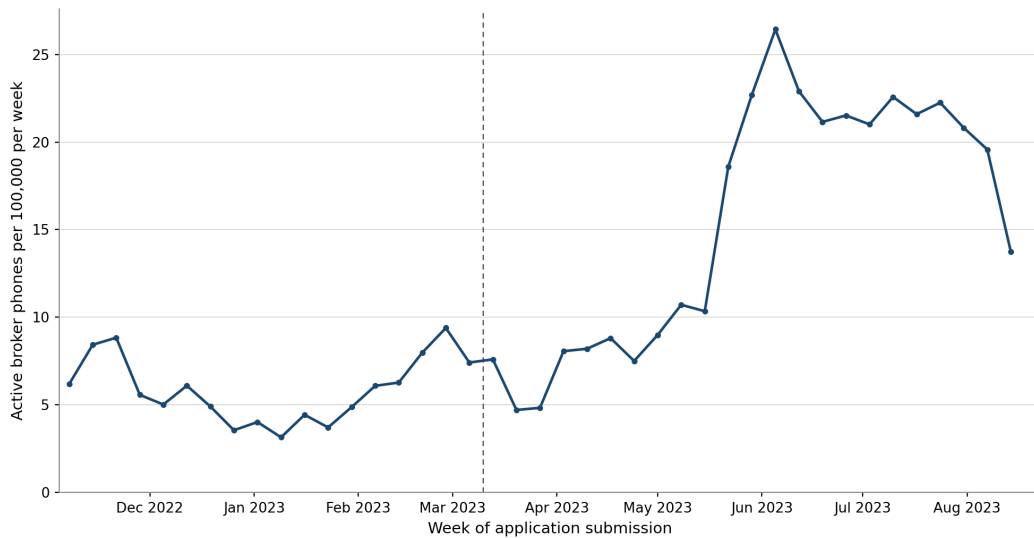


Figure 11. Active brokers per 100,000 populations



6. Instrument Variable Regression Results

Throughout this section, we discuss magnitudes for an increase of 10 frontline staff per 100,000 residents in a district as a benchmark. This is because the minimum that any district in our sample has seen after the statewide hiring shock is 10 staff per 100,000 population in a day. The average is much higher: districts saw an average of 120 staff per 100,000 population after the hiring shock

relative to 75 before the shock. For each outcome, we report percentage change as:

$$100 \times \frac{10\hat{\beta}}{\bar{Y}_{\text{pre}}}, \quad (6)$$

where $\hat{\beta}$ is the IV coefficient and $\bar{Y}_{\text{pre}}$ is the pre-shock mean outcome at the district-date level.

6.1. Public service delivery outcomes

The first result is that the frontline expansion increased application volume. Table 2 reports positive and statistically significant effects on applications per capita. The IV estimate implies that adding 10 frontline staff per 100,000 residents increases applications per capita by 1.88 relative to a pre-shock mean of 11.45, or about 16.4 percent. The OLS estimates are slightly smaller than our IV estimates. It makes intuitive sense for the OLS estimates to be biased towards zero since actual local staffing may have been targeted at districts. On the other hand, we also do not expect the estimates to differ by a large margin either because frontline operators make up only a part of the production process that determines application volume. It is also the stage which does not directly impact throughput since certificate issuance authority lies with approvers.

The IV estimate for average processing time of applications shows a decline of 0.43 days relative to a pre-shock mean of 14.22 days, or about 3.0 percent. This is informative because the increase in applications could mechanically create congestion. The fact that average processing time declines while volume rises implies that the direct capacity effect of additional frontline operators dominates any slowdown that may have happened due to higher applications being submitted in response to the additional frontline operators.

Table 2. Effects of frontline staffing on application volume and processing time

	Applications per capita		Processing time	
	OLS	IV	OLS	IV
Frontline staff per 100k	0.154*** (0.019)	0.188*** (0.025)	−0.035** (0.011)	−0.043** (0.013)
Observations	6,569	6,569	6,569	6,569
Pre-shock mean	11.45	11.45	14.22	14.22
Kleibergen-Paap F-stat		15.55		15.55
District FE	Yes	Yes	Yes	Yes
Day-of-week FE	Yes	Yes	Yes	Yes

Notes: Standard errors in parentheses. Staff per capita is measured as frontline staff per 100,000 residents. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Applications by submission channel. Table 3 reports the staffing response separately for brokered and direct applications. The IV coefficients imply that an increase of 10 frontline workers per 100,000 residents raises brokered applications by approximately 0.55 and direct applications by 1.46 per 100,000 residents. Both estimates are positive and statistically significant.

Table 3. Effects of frontline staffing on brokered and direct applications

	Brokered applications		Direct applications	
	OLS	IV	OLS	IV
Frontline staff per 100k	0.046*** (0.007)	0.055*** (0.009)	0.116*** (0.014)	0.146*** (0.018)
Observations	6,408	6,408	6,408	6,408
Pre-shock mean	2.282	2.282	8.693	8.693
Kleibergen–Paap F-stat		15.933		15.933
District FE	Yes	Yes	Yes	Yes
Day-of-week FE	Yes	Yes	Yes	Yes

Notes: Standard errors in parentheses. Frontline staffing and application volumes are measured per 100,000 residents. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

6.2. Broker market outcomes

Mean processing time. Table 4 shows that the difference in mean processing time of direct and brokered applications narrowed. The IV estimate implies that adding 10 frontline staff per 100,000 residents reduces the broker speed advantage by 0.41 days, or 11.6 percent of the pre-shock gap. This is important for the model because the convergence is not primarily driven by brokered applications becoming slower. Brokered processing time falls by an imprecise 0.10 days (about 0.9 percent of its pre-shock mean), while non-brokered processing time falls by 0.51 days, about 3.4 percent of its pre-shock mean. It is the direct application line that catches up.

Broker market size. Despite the smaller speed gap, the broker market expanded. Table 5 shows increases in the share of total applications submitted via brokers as well as the total number of active brokers, suggesting new broker entry. For an increase of 10 frontline staff per 100,000 residents, broker share of applications rises by 0.65 points relative to a pre-shock mean of 19.65, a 3.3 percent increase. Active brokers rise by 0.92 relative to a pre-shock mean of 10.84, which is an 8.5 percent increase.

Table 4. Broker speed differential and decomposition

	Speed differential IV	Brokered time IV	Non-brokered time IV
Frontline staff per 100k	−0.041*** (0.009)	−0.010 (0.008)	−0.051** (0.014)
Observations	6,438	6,439	6,568
Pre-shock mean	3.54	11.35	14.95
Kleibergen-Paap F-stat	15.55	15.55	15.55
District FE	Yes	Yes	Yes
Day-of-week FE	Yes	Yes	Yes

Notes: Standard errors in parentheses. Staff per capita is measured as frontline staff per 100,000 residents. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 5. Broker market responses

	Broker share IV	Active brokers IV
Frontline staff per 100k	0.065*** (0.019)	0.092*** (0.010)
Observations	6,408	6,546
Pre-shock mean	19.65	10.84
Kleibergen-Paap F-stat	15.93	15.45
District FE	Yes	Yes
Day-of-week FE	Yes	Yes

Notes: Standard errors in parentheses. Active brokers are measured per 100,000 residents. Staff per capita is measured as frontline staff per 100,000 residents. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

6.3. Elasticities and economic interpretation

The pre-reform mean of frontline staffing in the main tables is

$$\bar{F}_{\text{pre}} = 75$$

workers per 100,000 population. Since the main regressions are estimated in levels, we convert each estimated coefficient into an elasticity using

$$\varepsilon_{Y,F} = \hat{\beta}_Y \frac{\bar{F}_{\text{pre}}}{\bar{Y}_{\text{pre}}}$$

Elasticities for our outcomes of interest are outlined below in Table 6³.

³For comparison, the accounting identities $Q_B = S_B Q$ and $Q_D = (1 - S_B)Q$ imply $\varepsilon_{Q_B,F} = \varepsilon_{Q,F} + \varepsilon_{S_B,F}$ and $\varepsilon_{Q_D,F} = \varepsilon_{Q,F} - [S_B/(1 - S_B)]\varepsilon_{S_B,F}$. Applying these expressions to the total-application elasticity of 1.23, broker-share elasticity of 0.25, and broker share of 0.1965 gives 1.48 and 1.17, respectively.

Table 6. Elasticities evaluated at pre-reform means

Outcome	IV coefficient	Pre-reform mean	Elasticity
Total applications	0.188	11.45	1.23
Broker share	0.065	19.65	0.25
Active brokers per 100 applications	0.092	10.84	0.64
Mean processing time	-0.043	14.22	-0.23
Brokered processing time	-0.010	11.35	-0.07
Direct processing time	-0.051	14.95	-0.26
Direct-minus-broker time gap	-0.041	3.54	-0.87
<i>Application volumes estimated separately</i>			
Brokered applications	0.055	2.282	1.70
Direct applications	0.146	8.693	1.09

Brokered share of marginal applications. The channel-specific IV coefficients also show how the additional applications are divided between the two routes. Let $\hat{\beta}_B$ and $\hat{\beta}_D$ denote the staffing coefficients for brokered and direct applications. Since total applications equal the sum of the two channels, using the same specification, sample, and weights gives an implied total response of $\hat{\beta}_B + \hat{\beta}_D$. The brokered share of marginal applications is therefore

$$\hat{m}_B = \frac{\hat{\beta}_B}{\hat{\beta}_B + \hat{\beta}_D} = \frac{0.055}{0.055 + 0.146} \simeq 0.274.$$

Thus, about 27.4 percent of the additional applications implied by these estimates are brokered. The corresponding baseline share is the ratio of pre-reform mean brokered applications to mean total applications in the same sample:

$$\bar{s}_{\text{pre}} = \frac{\bar{Q}_{B,\text{pre}}}{\bar{Q}_{B,\text{pre}} + \bar{Q}_{D,\text{pre}}} = \frac{2.282}{2.282 + 8.693} \simeq 0.208.$$

Brokered applications therefore account for about 20.8 percent of baseline application volume, compared with 27.4 percent of marginal applications. The difference is approximately 6.6 percentage points, or a 31.6 percent higher share relative to baseline:

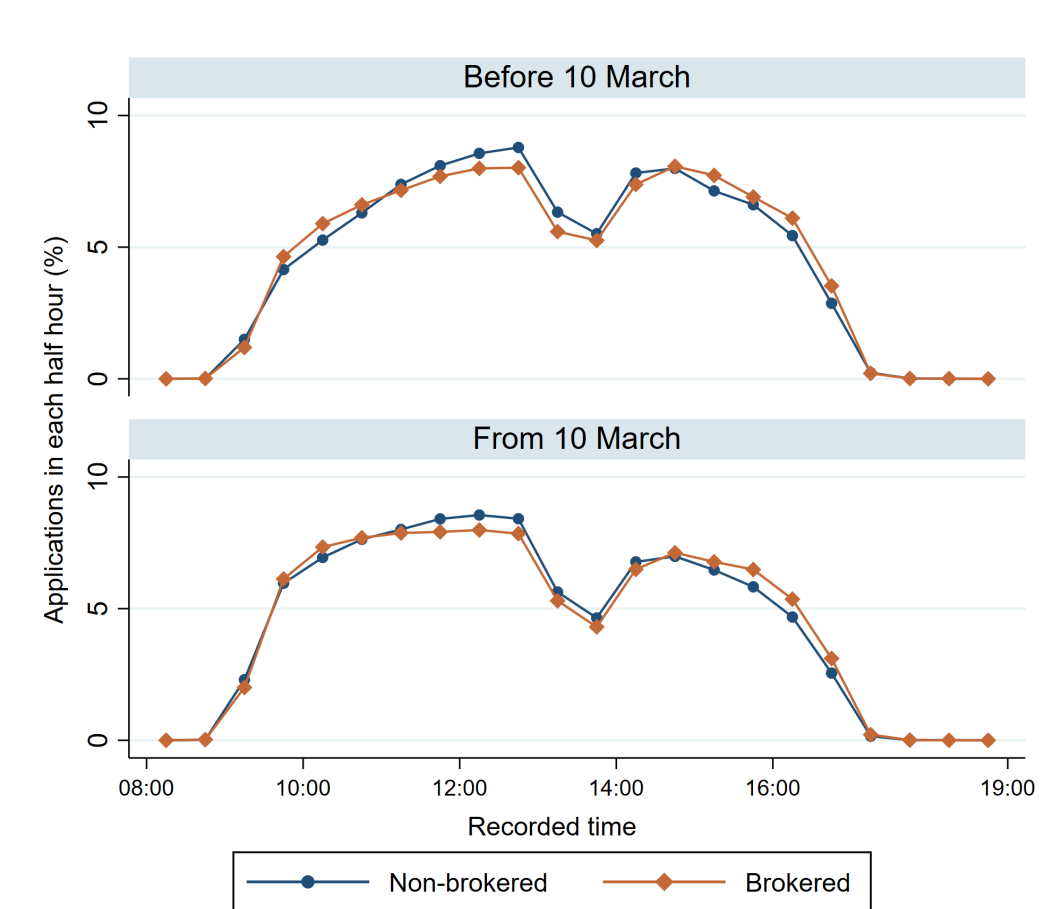
$$100 \left(\frac{\hat{m}_B}{\bar{s}_{\text{pre}}} - 1 \right) \simeq 31.6\%.$$

6.4. Where does the remaining broker advantage sit?

Finally we try to understand whether some other observed broker advantage persists in the system despite the converging mean processing time gap. We consider the idea of early approvals, i.e., probability of getting approved before the statutory deadline. In theory frontline operators can do

this by reserving earlier tokens in the SK counters for brokers and their clients⁴. This advantage can be provided by frontline operators if they systematically let brokers submit applications early. This is the only margin on which operators can ensure brokered applications enter the system before direct applications and are therefore processed earlier than direct ones. We show that this is unlikely to be true in Figure 12. Both brokered and direct applications enter the system roughly at the same time throughout the day. This means that tokens for morning slots were unlikely to be distributed systematically to brokers.

Figure 12. Time of application submission



Another way to obtain early approvals is when applications reach approvers. We do not observe variation at the approver stage. We also do not observe when an application reaches the approvers after being entered into the system by a frontline operator. Therefore this exercise is purely suggestive. We distinguish between discretionary and routine services based on their application requirements. Discretionary services require at least one signature from a senior

⁴During fieldwork we found this to be sometimes common although not widespread.

government official, often the approver, whereas routine services typically rely on existing documents, such as proof of address or income, without requiring additional authorization from a senior official. Help with paperwork and submission could benefit applicants in both groups. Privileged access to approvers, however, could provide an additional advantage in services where officials have greater discretion over approval. We compare how the broker–direct gap in the probability of approval by the statutory deadline changes across these two groups following the frontline expansion. A broker advantage that persists in discretionary services but diminishes in routine services would be consistent with privileged access. However, we cannot rule out differences in documentation requirements or applicant composition can also generate this pattern. In our sample, most health department services and income certificates are routine, whereas caste certificates, non-availability of birth/death certificates, widow pension certificates are discretionary.

For application i in service s , let r_i denote the approval date relative to that service’s statutory deadline and define

$$\text{Early}_i \equiv \mathbf{1}\{r_i \leq 0\}. \quad (7)$$

Thus, Early_i equals one when the application is approved on or before its service-specific statutory deadline. Let B_i indicate a broker-associated application, let H_s indicate a high-discretion service, and let Post_t indicate the period after the frontline staffing expansion. For discretion group $g \in \{L, H\}$ and period $p \in \{\text{pre}, \text{post}\}$, define the broker early-approval gap as

$$G_{g,p} \equiv \Pr(\text{Early} = 1 \mid B = 1, g, p) - \Pr(\text{Early} = 1 \mid B = 0, g, p). \quad (8)$$

Figure 13 reports the four gaps taking into account service composition. Before the reform, broker advantage is similar across the two groups: $G_{H,\text{pre}} = 2.16$ percentage points and $G_{L,\text{pre}} = 2.43$ percentage points. After the reform, the low-discretion gap falls to 0.97 percentage points, while the high-discretion gap remains essentially unchanged at 2.17 percentage points. The low-discretion broker advantage therefore falls by about 1.46 percentage points, whereas the high-discretion advantage changes by only 0.01 percentage points. However, this is imprecisely estimated and we only consider the estimates as being suggestive.

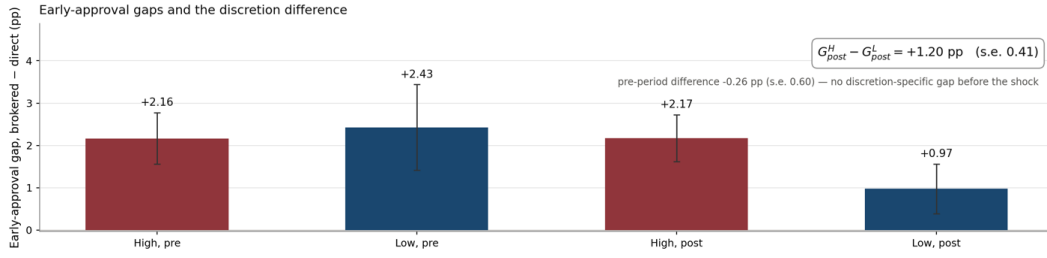
The primary estimand is the change in the high-minus-low broker gap after the reform:

$$\begin{aligned} \Delta^{DDD} &\equiv (G_{H,\text{post}} - G_{H,\text{pre}}) - (G_{L,\text{post}} - G_{L,\text{pre}}) \\ &\approx (2.17 - 2.16) - (0.97 - 2.43) = 1.47 \text{ percentage points.} \end{aligned} \quad (9)$$

Table 7. Decomposing the broker early-approval advantage

	Pre-reform	Post-reform	Change
Low-discretion services	2.43	0.97	−1.46
High-discretion services	2.16	2.17	+0.01
High minus low	−0.26	+1.20	+1.47

Notes: We report the percentage-point difference in the probability of early approval between brokered and direct applications. The final column is the post-minus-pre change. The lower-right entry is the DDD estimand from equation (9).

Figure 13. Broker early-approval advantage by service discretion

We estimate a linear-probability specification by:

$$\begin{aligned} \text{Early}_{isdt} = & \alpha_s + \mu_d + \lambda_t + \beta_B B_i + \beta_{BP} (B_i \times \text{Post}_t) + \beta_{BH} (B_i \times H_s) \\ & + \beta_{HP} (H_s \times \text{Post}_t) + \beta_{DDD} (B_i \times H_s \times \text{Post}_t) + X_i' \gamma + \varepsilon_{isdt}, \end{aligned} \quad (10)$$

where α_s , μ_d , and λ_t are service, district, and submission-time fixed effects. The interaction $H_s \times \text{Post}_t$ absorbs differential post-reform changes in high-discretion services that are common to direct and brokered applications. In this specification, β_{DDD} is the DDD estimand in (9).

Interpretation. The post-reform high-minus-low difference is

$$G_{H,\text{post}} - G_{L,\text{post}} = 2.17 - 0.97 = 1.20 \text{ percentage points.} \quad (11)$$

If low-discretion services provide a valid benchmark for the generic broker advantage, including applicant selection, application preparation, and other route differences that do not depend on approval discretion, then the 0.97-point low-discretion gap is a residual component and the additional 1.20 points in high-discretion services are a discretion-linked excess advantage. On this descriptive decomposition, roughly 55 percent of the post-reform high-discretion broker gap is associated with the discretion-linked component.

Besides the imprecision, we also avoid causal interpretation because high- and low-discretion

services may differ along other margins. Instead, we consider this result as helping us locate a potential residual broker advantage. That is, after frontline capacity expands, the remaining broker advantage may sit in early approval concentrated in services where downstream officials retain greater discretion.

7. Benchmark models of the broker market

The empirical results present a simple puzzle. Additional application-stage workers increase applications through both channels, with a larger response through brokers. More brokers also become active. Yet direct applications become faster, reducing the observed speed advantage of brokerage. We use three benchmarks to ask which services can account for these responses. Under the restrictions below, pure substitution predicts a falling broker share while pure approval priority predicts a widening observed time advantage. Complementarity allows brokerage to expand despite a smaller time advantage.

Pure substitution benchmark. The common belief of brokers substituting for scarce public capacity is our first benchmark model. While we show a complete proof below, we can also intuitively see how a pure substitution framework will not be able to predict all empirical facts jointly. Most importantly, if the government becomes easier to access, and there is an increase in total applications, the number of brokered applications might rise. But this model does not predict a rise in the *share* of brokered applications coinciding with falling observed speed advantage. Admittedly, citizens do not only care about the time taken after an application is submitted, they also care about the time taken to reach submission. The predictions would be affected if time taken to reach submission (which is unobserved in our administrative data) varies systematically by brokers and direct applicants. Here we rely on our observation that the time of submission does not vary greatly across direct and brokered applicants. This means that on average, brokered applications do not reach the counters earlier than direct applicants over the day. But there is a possibility of brokers having greater expertise and can generally take less time to prepare an application package – we would not be able to observe this in our data⁵. For the pure substitution benchmark, we assume away any additional advantage brokers might have that frontline workers do not, and hence they are pure substitutes.

⁵We are collecting mystery shopper survey data on prices charged by brokers and the nature of services they provide with documentation. 30 brokers have been surveyed so far. We will update the draft with results as soon as we are done.

Pure approval-stage priority benchmark. Under a framework of pure priority at approver-stage, the main job a broker has is precisely to be ahead of a direct application. This means even if total applications rise, the brokered queue becomes longer, and the value to jumping queue increases, brokers only survive as long as their speed advantage remains. This model cannot jointly predict a rise in broker share along with a falling broker speed advantage.

Complementarities benchmark. Lastly, we consider how brokers might actually work with frontline operators (since they too need to submit the application at the SK). Higher number of frontline workers can make brokers more productive by reducing the effort cost a broker incurs to reach an operator. This scenario is realistic and is also capable of jointly rationalizing all our empirical facts.

7.1. A common environment

Consider one district and one service. A citizen receives value $V > 0$ from obtaining the service and has a time cost θ . Potential applicants have constant density $\lambda > 0$ on $[0, \bar{\theta}]$. Citizens can apply directly, hire a broker at fee P , or not apply. Let Q_D and Q_B denote the flows of direct and brokered applications, $Q = Q_D + Q_B$ their total, and $S_B = Q_B/Q$ the brokered share. The value of the service and the distribution of citizens are held fixed when staffing changes.

Applications require frontline work and approval. The reform increases frontline staff-time F while holding approval capacity μ fixed. Feasible stationary flows satisfy $L = \phi_D Q_D + \phi_B Q_B < F$ and $Q < \mu$, where ϕ_j is public frontline work per application. Brokered clients' use of public workers is included in L .

Citizens care about the full application process, while the records begin at formal submission. Write R_j for preparation time and T_j for recorded processing time, where $j \in \{D, B\}$. Define

$$\tau = R_B + T_B, \quad H = (R_D + T_D) - (R_B + T_B) > 0, \quad \Delta = T_D - T_B. \quad (12)$$

Here τ is the full broker-route time, H is the full time advantage of brokerage, and Δ is its recorded advantage. A fall in Δ does not establish how preparation time or H changes.

The utilities of direct application, brokerage, and nonparticipation are

$$U_D = V - \theta(\tau + H), \quad U_B = V - P - \theta\tau, \quad U_0 = 0. \quad (13)$$

In an interior allocation (i.e. one where a broker market exists), citizens with low time costs apply directly, those with intermediate time costs hire brokers, and those with the highest time costs do not apply. The resulting demands are

$$Q_D = \lambda \frac{P}{H}, \quad Q_B = \lambda \left(\frac{V - P}{\tau} - \frac{P}{H} \right), \quad Q = \lambda \frac{V - P}{\tau}. \quad (14)$$

Participation therefore responds to both application time and broker fees.

Brokers have marginal cost c and fixed operating cost $K > 0$. Quantity competition and free entry determine fees, applications per broker, and active brokerage. Appendix C.1 derives this equilibrium. All comparative statics include congestion from changes in participation and route choice.

7.2. Substitution: public assistance replaces broker assistance

In this benchmark, brokers help prepare and correct applications but receive no approval priority. Their work is identical to frontline operators. They only need to additionally engage with frontline staff to submit the application package. Direct applications require more frontline handling. Increasing frontline staff leading to faster frontline interactions thus narrow the recorded gap. Brokers' marginal and operating costs remain fixed in this benchmark. Two demand forces compete: a smaller full broker advantage encourages direct application, while a shorter common application process encourages participation. A substitution prediction must account for both forces, as well as the brokers' endogenous fee.

Let $x = -d \log \tau / d \log F$ denote the proportional improvement in the full broker route and $y = -d \log H / d \log F$ the proportional reduction in its full advantage. The crowd-out benchmark maintains $y > 0$ and restricts common access gains to satisfy

$$\frac{x}{y} \leq \frac{\tau}{2\tau + H}. \quad (15)$$

Prediction 1. With constant broker costs, an interior equilibrium, and (15), additional frontline capacity reduces the brokered share. This benchmark cannot explain the observed increase in S_B .

Public assistance reduces the reason to hire a broker faster than easier access creates new demand for brokerage. Fees and participation adjust but do not reverse crowd-out under the bound. This restriction concerns full application time: the recorded gap cannot establish it without preparation-time assumptions. We consider this more seriously in the complementarities benchmark. Sufficiently large improvements in unobserved preparation can produce brokerage

growth at constant costs outside this benchmark. The rejection therefore applies to the substitution model with the assumptions we state. Appendix C.4 gives the proof.

7.3. Pure priority: brokers obtain earlier approval

In the second benchmark, brokers provide no productive assistance. Both routes require identical frontline work, and brokerage only purchases priority at the approval stage. The approval facility completes the application already in service before selecting a waiting brokered application ahead of a direct one. Approval capacity, service requirements, and this priority rule remain unchanged after the staffing increase. The appendix formalizes the facility as a stationary queue with Poisson arrivals and identical exponential service requirements.

The common frontline process cancels from the recorded time gap. Under the specified priority rule, the approval advantage satisfies

$$\Delta = g(Q, Q_B; \mu), \quad g_Q > 0, \quad g_{Q_B} > 0. \quad (16)$$

Prediction 2. At a fixed approval facility, an increase in both total and brokered applications widens the recorded broker advantage. Pure priority therefore cannot explain expanding application flows and a narrowing recorded gap.

More applications raise congestion, and more priority applications increase the delay imposed on direct applicants. This result allows spare approval capacity and does not depend on how brokers or officials price priority. A faster common frontline process could still lower direct recorded time; it cannot reverse the gap prediction because both routes receive the same frontline improvement. The contradiction concerns priority as the sole broker service under the stated queueing assumptions. Appendix C.5 derives the result.

7.4. Productive complementarity

The third benchmark retains the assistance model's public technology and citizen choices. Its distinguishing feature is broker production. Brokers repeatedly interact with public workers to prepare, correct, and submit applications. When these interactions take less time, a broker can serve a client with less private effort. Let ω_p measure the time required for pre-submission interactions and e_m a broker's effort. A simple technology is

$$q_m = e_m \left(\frac{\bar{\xi}}{\omega_p} \right)^\psi, \quad c = c_0 + \bar{c} \left(\frac{\omega_p}{\bar{\xi}} \right)^\psi, \quad \bar{\xi}, \bar{c}, \psi > 0, \quad c_0 \geq 0 \quad (17)$$

The cost expression follows when private effort costs $\bar{c}$ per unit and other costs are c_0 per client. Faster interactions lower broker cost; brokered clients' public work remains in the resource constraint. Holding c fixed recovers the substitution cost specification. Both benchmarks use the same preparation and recorded-time technologies.

Lower costs support a lower fee and encourage participation, counteracting the declining broker advantage. Direct applications can increase at the same time as relative time savings favor the direct route and easier access brings new citizens into the market.

Let $k = -d \log P / d \log F$ be the proportional fall in the broker fee. Direct applications and the brokered share both increase when

$$\max \left\{ 0, \frac{V-P}{V}(y-x) \right\} < k < y, \quad x \geq 0, \quad y > 0 \quad (18)$$

The lower bound makes the fee decline strong enough to raise broker share; the upper bound leaves direct application sufficiently more attractive to raise direct volume. Total applications then rise as well. Broker entry requires brokered volume to grow faster than applications per broker, which also adjust in equilibrium.

The time response imposes a separate requirement. In the assistance models, recorded time is $T_j = \ell_j \omega_s + d(Q; \mu)$, where $\ell_D > \ell_B > 0$ is post-submission frontline work, ω_s its elapsed time per unit, and $d(Q; \mu)$ common approval time. Faster frontline interactions narrow the gap. They also reduce direct recorded time if the frontline saving outweighs the approval congestion induced by additional applications. Brokered recorded time can change little when its smaller frontline saving approximately offsets the same increase in approval delay.

Prediction 3. If the broker-cost reduction supports (18), brokered volume grows faster than individual broker scale, and frontline time savings outweigh the increase in direct applicants' approval delay, productive complementarity generates the joint response: both application routes expand, broker share and entry rise, direct processing time falls, and the recorded gap narrows. Appendix C.6 gives the exact conditions and their derivation, including both public capacity constraints.

Public capacity can therefore reduce the assistance advantage of brokerage while increasing brokers' productivity. Under the stated conditions, that cost channel reconciles the joint responses that the competing benchmarks cannot. Brokers may also retain an approval-stage advantage; the

priority result concerns an explanation based on priority alone.

8. Discussion of citizen welfare

Citizen welfare depends on the full cost of obtaining a service. For a citizen with time cost θ , welfare in the model is

$$\mathcal{U}(\theta; F) = \max \{0, V - \theta(\tau + H), V - P - \theta\tau\}.$$

Under the complementarity conditions derived above, a marginal staffing increase lowers full direct application time, weakly lowers full brokered application time, and lowers the broker fee. Every citizen is therefore weakly better off. The previous route remains available at weakly lower cost, and citizens can change routes or begin applying. This welfare gain is compatible with an increase in broker use. The substitution benchmark can likewise improve welfare when full application times fall, despite predicting crowd-out.

Approval priority introduces distributional tradeoffs. Earlier approval for brokered applications can increase waiting time for direct applicants. At a fixed approval facility, priority creates no additional processing capacity, although allocating shorter waits to citizens with higher time costs can reduce the total cost of waiting. Broker fees also reduce citizen surplus, but the payments themselves are transfers to brokers. An assessment of net social welfare must additionally account for the type of applicants who are unable to pay for broker services, real broker effort, operating costs, and the cost of public staffing.

The administrative evidence documents faster recorded processing for direct applications, which captures one component of the cost of access. It does not recover preparation-time or fee changes, so it is difficult to establish the model’s full welfare conditions or quantify the change in citizen surplus. The welfare interpretation therefore remains conditional on these unobserved responses.

9. Conclusion

Better public services need not reduce citizens’ use of private brokers. We study a staffing reform in a large Indian state that added workers to help citizens prepare and submit applications for government documents. These documents are often required to access public benefits and services. The reform expanded help at the application stage while leaving approval staffing essentially unchanged. Applications increase through both direct and brokered channels, but the increase

is larger through brokers. The number of active brokers also increases. Direct applications are processed more quickly, narrowing the observed broker advantage. Brokerage therefore expands even as direct access improves.

In this paper, we see an instance of how public workers can make brokers cheaper to operate. Citizens need less private help when public staff can check documents, answer questions, and correct mistakes. Brokers also depend on these workers to prepare and submit applications for their clients. Easier access to public workers can therefore reduce the effort needed to serve each client. These lower costs translate to lower broker prices (as predicted in the model). This makes private assistance cheaper, encourage participation, and supports broker entry. Under the maintained restrictions, the substitution benchmark predicts a falling broker share. Priority alone predicts constant or higher processing-time gaps between direct and brokered applications as application flows rise at a fixed approval facility.

While we establish that complementarity can rationalize all of our empirics, we also show that better assistance at the application stage need not remove broker advantages at the approval stage. The broker advantage in approval by the statutory deadline falls substantially in lower-discretion services. It remains essentially unchanged in higher-discretion services. This pattern is consistent with brokers retaining an advantage where officials have more scope to influence outcomes. We do not have enough evidence from the administrative data to establish how that advantage is obtained⁶. Help with paperwork and repeated visits can reduce real costs. Access to priority can change who receives scarce official attention without increasing approval capacity. The expansion of brokerage therefore need not imply that every service brokers provide creates additional surplus.

The broader point of the paper is that the cost of accessing the state depends on both public services and the private markets that form around them. Broker prevalence alone does not show how well/poorly the state serves citizens. The welfare consequences depend on what brokers provide, what citizens pay in money and time, and the distributional consequences of such an unregulated market. Administrative burdens therefore can create private markets for overcoming them, and those markets respond to changes in the state. Building state capacity therefore requires understanding both how government services are produced and how private markets adjust around them.

⁶The mystery shoppers in our ongoing fieldwork are currently trying to ask this to brokers

A. Results with a post-reform indicator

Table 8 reports OLS and IV estimates that include a post-reform indicator alongside district fixed effects. The indicator equals one from March 10, 2023 onward. It allows a common shift in outcomes at the reform date that is separate from the estimated staffing response.

Application volume and active brokerage continue to increase with staffing. The IV coefficients are 0.220 for total applications per capita, 0.050 for brokered applications per capita, and 0.170 for direct applications per capita. Active brokers per capita also increase, with an IV coefficient of 0.034. All four estimates are positive and statistically significant, as are their OLS counterparts. The expansion in broker-mediated applications therefore continues to coexist with an expansion in direct applications.

Processing times retain the main pattern. For an increase of ten frontline workers per 100,000 residents, the IV estimates imply a decline of 0.524 days in mean processing time. Direct processing time falls by 0.63 days, while brokered processing time falls by an imprecisely estimated 0.14 days. The direct-minus-broker gap narrows by 0.49 days. Thus, the narrowing gap continues to reflect faster direct processing rather than slower brokered processing. The estimated effect on the delay rate is small and statistically indistinguishable from zero.

The broker-share result is less robust. Its OLS coefficient remains positive at 0.032 (standard error 0.013), but its IV coefficient is close to zero and slightly negative at -0.005 (standard error 0.039). The IV estimate therefore differs from the positive baseline result in both its point estimate and its precision. This specification supports an increase in broker activity alongside improved direct processing, but does not establish that brokered applications expand disproportionately. The reported first-stage statistics are approximately 11.4 across outcomes.

Table 8. Staffing results with a post-reform indicator

Outcome	OLS	IV	Pre-reform mean	N	KP F (IV)
<i>A. Applications and brokerage</i>					
Total applications per capita	0.141*** (0.024)	0.220*** (0.044)	10.975	6,408	11.404
Brokered applications per capita	0.036*** (0.008)	0.050** (0.014)	2.282	6,408	11.404
Direct applications per capita	0.105*** (0.018)	0.170*** (0.032)	8.693	6,408	11.404
Broker share (percentage points)	0.032* (0.013)	−0.005 (0.039)	19.654	6,408	11.404
Active brokers per capita	0.024*** (0.005)	0.034*** (0.007)	1.082	6,408	11.404
<i>B. Processing times and delays</i>					
Mean processing time (days)	−0.0348** (0.0110)	−0.0524* (0.0197)	14.2939	6,408	11.4041
Delay rate (share)	−0.0004 (0.0005)	−0.0003 (0.0007)	0.3445	6,408	11.4041
Direct-minus-broker processing time (days)	−0.034*** (0.006)	−0.049** (0.015)	3.741	6,266	11.353
Brokered processing time (days)	−0.007 (0.010)	−0.014 (0.010)	11.122	6,266	11.353
Direct processing time (days)	−0.041** (0.012)	−0.063** (0.021)	14.863	6,266	11.353
District fixed effects	Yes	Yes			
Post-reform indicator	Yes	Yes			

Notes: Each row reports the coefficient on frontline staffing from separate OLS and IV regressions. Staffing is measured as frontline workers per 100,000 residents. Standard errors are in parentheses. All specifications include district fixed effects and an indicator equal to one from March 10, 2023 onward. Application and active-broker outcomes are reported per capita; broker share is measured in percentage points and the delay rate as a fraction. Processing times and the direct-minus-broker gap are measured in days. *N* is the number of district-date observations; the sample size and reported pre-reform mean are the same for OLS and IV within each row. KP *F* reports the Kleibergen–Paap first-stage statistic for the IV regression. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

B. Robustness

Figure 14. District exposure and pre-shock service delivery outcomes

District exposure does not predict pre-period public service delivery outcomes

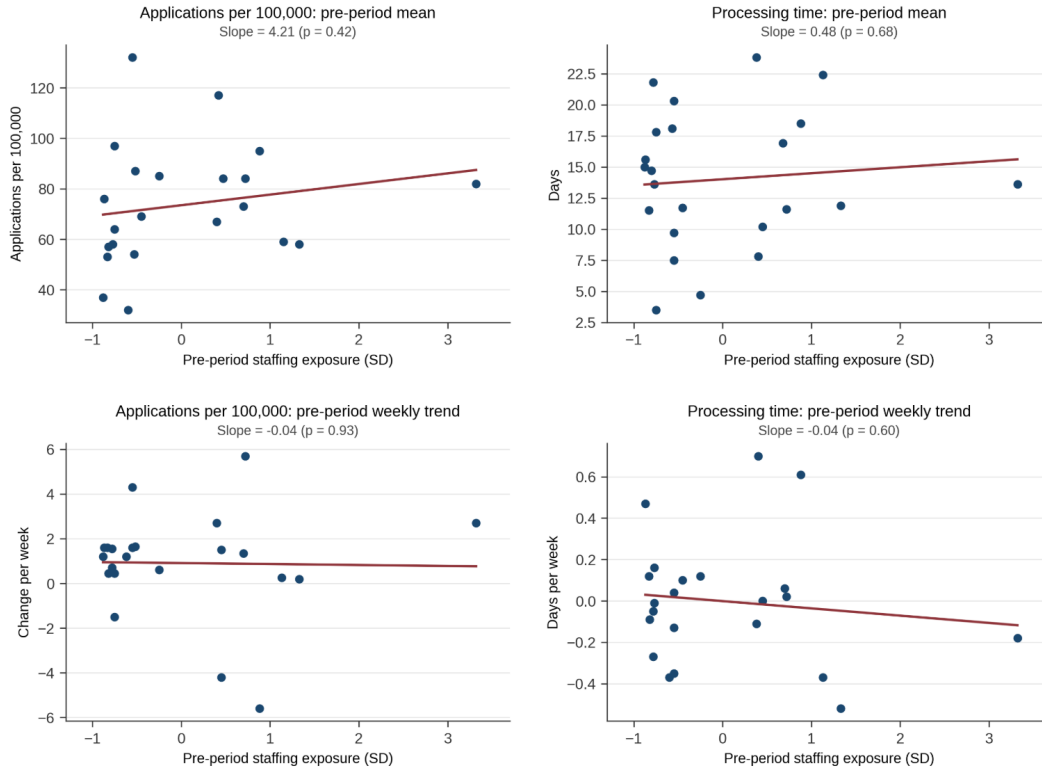
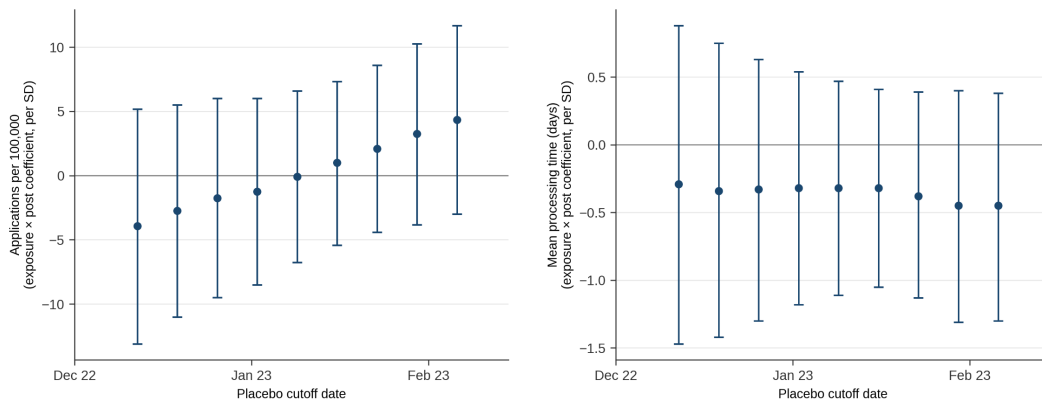


Figure 15. Placebo reform dates in pre-reform period for service delivery outcomes



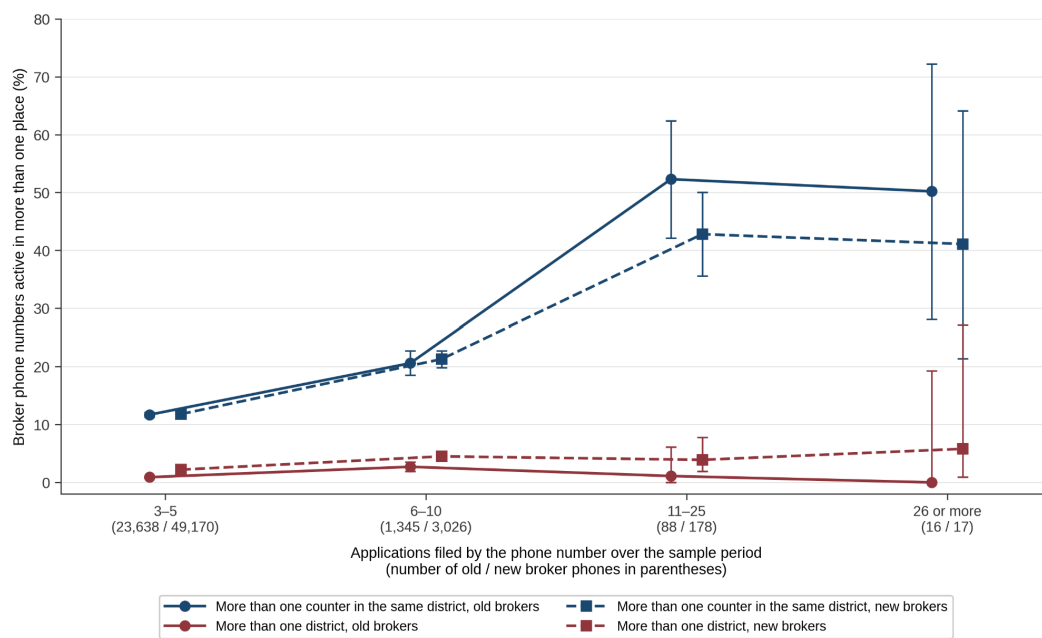
Notes: Each point is the coefficient on exposure $\times$ post from a separate regression that uses pre-reform data only. “Post” equals one for dates on or after the placebo cutoff shown on the horizontal axis. Placebo cutoffs are set weekly from Dec 2022 to Feb 2023. Exposure is district pre-reform frontline staff per 100,000 residents, standardized to mean zero and unit standard deviation. All regressions include district and day-of-week fixed effects. Bars show 95% confidence intervals based on standard errors clustered by district and date. The left panel reports applications per 100,000 residents and the right panel reports mean processing time in days.

Table 9. Broker-threshold sensitivity check

Minimum applications per phone	3	4	6	8	10
All applications per 100,000	0.188 (0.025)	0.188 (0.025)	0.188 (0.025)	0.188 (0.025)	0.188 (0.025)
District-date observations	6,569	6,569	6,569	6,569	6,569
Mean processing time (days)	-0.043 (0.013)	-0.043 (0.013)	-0.043 (0.013)	-0.043 (0.013)	-0.043 (0.013)
District-date observations	6,569	6,569	6,569	6,569	6,569
Direct minus broker time (days)	-0.041 (0.009)	-0.033 (0.007)	-0.025 (0.005)	-0.006 (0.013)	-0.002 (0.021)
District-date observations	6,438	6,149	4,694	2,980	2,063
Broker processing time (days)	-0.010 (0.008)	-0.010 (0.008)	-0.016 (0.012)	-0.040 (0.024)	-0.052 (0.031)
District-date observations	6,439	6,149	4,694	2,980	2,063
Direct processing time (days)	-0.051 (0.014)	-0.046 (0.013)	-0.044 (0.013)	-0.043 (0.013)	-0.043 (0.013)
District-date observations	6,568	6,569	6,569	6,569	6,569
Broker share (percentage points)	0.065 (0.019)	0.059 (0.010)	0.025 (0.006)	0.014 (0.004)	0.008 (0.003)
District-date observations	6,408	6,408	6,408	6,408	6,408
Active broker phones per 100 applications	0.092 (0.010)	0.066 (0.006)	0.027 (0.004)	0.014 (0.003)	0.009 (0.002)
District-date observations	6,546	6,546	6,546	6,546	6,546
District and day-of-week FE	Yes	Yes	Yes	Yes	Yes

Notes: Each cell is a separate 2SLS regression, with standard errors clustered by district and application date in parentheses. Coefficients are per additional staff member per 100,000 residents. We reproduce outcomes of Tables 5–7.

Figure 16. Broker mobile numbers active in more than one location, by application volume



Notes: Old brokers are numbers whose first application is before 10 March 2023. Numbers are grouped by the total number of applications they filed over the sample period. The blue series shows the percent of phones that filed applications at more than one SK counter within a single district. The red series shows the percent that filed applications in more than one district. Bars show 95% confidence intervals.

C. Deriving a model of citizen choice, broker competition, and state capacity

This appendix derives citizen demand, broker competition and entry, and the three benchmark predictions. The results concern a differentiable branch of feasible stationary equilibria. A dot denotes a total derivative with respect to $z = \log F$, holding approval capacity and all other primitives fixed. Thus $\dot{Y} = dY/d \log F$ and $\varepsilon_Y = \dot{Y}/Y = d \log Y / d \log F$.

C.1. Citizen choice and broker competition

Citizen cutoffs. From (13), direct application and brokerage provide equal utility at $\theta_1 = P/H$. Brokerage and nonparticipation provide equal utility at $\theta_2 = (V - P)/\tau$. Assume

$$0 < \theta_1 < \theta_2 < \bar{\theta}. \quad (19)$$

Below θ_1 , direct application dominates brokerage and gives positive utility. Between θ_1 and θ_2 , brokerage dominates direct application and gives positive utility. Above θ_2 , neither route gives positive utility. Integrating the constant density λ over these intervals gives (14). In particular,

$$1 - S_B = \frac{Q_D}{Q} = \frac{P\tau}{H(V - P)}. \quad (20)$$

Inverse demand. Holding the public delay forecasts (τ, H) fixed, define

$$h = \frac{1}{\tau} + \frac{1}{H}, \quad a = \frac{VH}{\tau + H}. \quad (21)$$

Citizen demand implies $Q_B = \lambda h(a - P)$ and hence $P = a - Q_B/(\lambda h)$. Broker m chooses q_m , taking its competitors' quantities and district-wide public delays as given, to maximize

$$\pi_m = (P - c)q_m - K. \quad (22)$$

The broker internalizes its effect on P . It does not internalize the effect of its individual quantity on district-wide congestion. This is a maintained approximation for a market with many brokers; aggregate delay forecasts must nevertheless be correct in equilibrium.

Symmetry and free entry. The first-order condition and the zero-net-profit condition are

$$P - c = \frac{q}{\lambda h}, \quad (P - c)q = K. \quad (23)$$

Writing $r = P - c$ for the markup, these imply

$$r = \sqrt{\frac{K}{\lambda h}}, \quad P = c + r, \quad q = \sqrt{\lambda h K}, \quad M = \frac{a - c}{r} - 1, \quad Q_B = Mq. \quad (24)$$

Conditional profit is strictly concave in own quantity because inverse demand has slope $-1/(\lambda h)$. Entry is treated as continuous, so M measures active broker units. An interior broker market requires $a > c + r$; at least one full broker requires $a - c \geq 2r$. The citizen-support restriction in (19) must also hold. At $a \leq c + r$, the interior formulas do not describe an active broker market. With an integer number of brokers, entry may occur in steps rather than through a smooth local response.

C.2. Application processing time and equilibrium

The common assistance environment. Let $b_j \geq 0$ and $\ell_j > 0$ denote frontline work before and after formal submission. Work per application is $\phi_j = b_j + \ell_j$, and

$$L = \phi_D Q_D + \phi_B Q_B < F, \quad Q_D + Q_B < \mu. \quad (25)$$

Let $\omega_p(F, L)$ and $\omega_s(F, L)$ be elapsed time per unit of frontline work before and after submission. Both are differentiable, increasing in workload, and decreasing in staffing when workload is held fixed. These are reduced-form interaction and delay technologies. They count elapsed time rather than additional staff-time allocations. The two clocks are

$$R_j = \bar{R}_j + b_j \omega_p(F, L), \quad T_j = \ell_j \omega_s(F, L) + d(Q; \mu), \quad d_Q > 0, \quad (26)$$

where $\bar{R}_j \geq 0$ is fixed preparation time and d includes approval waiting and service. The assistance benchmarks maintain $\ell_D > \ell_B$, $b_D \geq b_B$, and $\bar{R}_D \geq \bar{R}_B$. Consequently,

$$\begin{aligned} \tau &= \bar{R}_B + b_B \omega_p + \ell_B \omega_s + d(Q; \mu), \\ H &= (\bar{R}_D - \bar{R}_B) + (b_D - b_B) \omega_p + (\ell_D - \ell_B) \omega_s, \\ \Delta &= (\ell_D - \ell_B) \omega_s. \end{aligned} \quad (27)$$

Substitution and complementarity use the same specifications for both clocks and public work. Only the broker-cost technology differs. No preparation-time improvement is allowed exclusively in the complementarity benchmark.

Equilibrium consistency. For candidate flows (Q_D, Q_B) , compute L , the public delays, (τ, H) , c , and the fee in (24). Equilibrium requires the candidate flows to equal citizen demand in (14). Define these two excess-flow equations by

$$E_D = Q_D - \lambda P / H = 0, \quad E_B = Q_B - \lambda [(V - P) / \tau - P / H] = 0. \quad (28)$$

Assume the Jacobian $J = \partial(E_D, E_B)/\partial(Q_D, Q_B)$ is nonsingular at the equilibrium under consideration. The implicit-function theorem then gives the local response

$$\begin{pmatrix} \dot{Q}_D \\ \dot{Q}_B \end{pmatrix} = -J^{-1} \frac{\partial(E_D, E_B)}{\partial z}. \quad (29)$$

All time and fee responses used below are evaluated along this equilibrium branch. For example,

$$\dot{\omega}_v = F \frac{\partial \omega_v}{\partial F} + \frac{\partial \omega_v}{\partial L} \dot{L}, \quad \dot{L} = \phi_D \dot{Q}_D + \phi_B \dot{Q}_B, \quad v \in \{p, s\}. \quad (30)$$

More staffing need not lower equilibrium delays when induced workload is sufficiently large. This is why the positive result imposes conditions on the total delay responses, rather than just their partial derivatives. Strict inequalities in (25) and the interior conditions remain valid for sufficiently small staffing changes on a continuous equilibrium branch. The local results do not apply to an overloaded approval facility with $Q \geq \mu$.

C.3. Useful equilibrium derivatives

Define

$$x = -\frac{\dot{\tau}}{\tau}, \quad y = -\frac{\dot{H}}{H}, \quad k = -\frac{\dot{P}}{P}, \quad v = \frac{\dot{h}}{h} = \frac{H}{\tau + H}x + \frac{\tau}{\tau + H}y. \quad (31)$$

Because K and λ are fixed, the broker-market formulas give

$$\frac{\dot{r}}{r} = -\frac{v}{2}, \quad \frac{\dot{q}}{q} = \frac{v}{2}, \quad k = -\frac{\dot{c}}{P} + \frac{r}{2P}v. \quad (32)$$

Differentiating citizen demand and (20) yields

$$\begin{aligned} \varepsilon_{Q_D} &= y - k, \\ \varepsilon_Q &= x + \frac{P}{V - P}k, \\ \dot{S}_B &= (1 - S_B) \left[x + \frac{V}{V - P}k - y \right]. \end{aligned} \quad (33)$$

These identities include participation and the changing broker fee. Route volumes, total applications, and broker share are linked by accounting and are not independent restrictions.

C.4. Proof of the substitution prediction

With $\dot{c} = 0$, equation (32) gives $k = rv/(2P)$. Define

$$w = \frac{H}{\tau + H}, \quad \zeta = \frac{Vr}{2P(V - P)}. \quad (34)$$

The sign of $\dot{S}_B$ is the sign of

$$(1 + \zeta w)x - [1 - \zeta(1 - w)]y. \quad (35)$$

Since $c \geq 0$, $r \leq P$. An active broker market also implies $P < a = Vw$, so

$$\zeta < \frac{1}{2(1 - w)}. \quad (36)$$

In particular, $1 - \zeta(1 - w) > 0$. For $y > 0$, the exact threshold for crowd-out is therefore

$$\frac{x}{y} < \frac{1 - \zeta(1 - w)}{1 + \zeta w}. \quad (37)$$

The fraction on the right decreases strictly with ζ . Using its upper bound gives

$$\frac{1 - \zeta(1 - w)}{1 + \zeta w} > \frac{1 - w}{2 - w} = \frac{\tau}{2\tau + H}. \quad (38)$$

Hence the weak bound in (15) implies strict crowd-out: $\dot{S}_B < 0$. This proves Prediction 1. The result allows the markup and participation to adjust. It is a sufficient condition for a declining share, not a claim that brokered volume must fall under every improvement in public assistance.

The unobserved clock qualification follows from

$$H = (R_D - R_B) + \Delta, \quad \tau = R_B + T_B. \quad (39)$$

Recorded processing-time responses do not determine x/y without assumptions on R_D and R_B . If common access improves sufficiently relative to H , the expression in (35) can become positive with constant costs. The analytical rejection is therefore conditional on the full-time restriction, not established by the recorded gap alone.

C.5. Proof of the pure priority prediction

Let $B = Q_B$ and $D = Q_D$, so $Q = B + D$. Both routes require identical public work. At the approval facility, arrivals from each route are independent Poisson processes, service times are exponential with common mean $1/\mu$, and $Q < \mu$. Service is nonpreemptive: an application already in service is completed before the next one is chosen. Waiting brokered applications have strict priority, with first-in-first-out service within each route.

Deriving the waiting times. An arrival sees an expected residual service time, including the possibility that the facility is idle, of $W_0 = Q/\mu^2$. This equals utilization Q/μ times the exponential distribution's mean residual service time $1/\mu$. Let W_B and W_D denote mean waiting times before service. A brokered arrival waits for residual service and brokered applications already waiting.

Little's law gives

$$W_B = W_0 + \frac{B}{\mu} W_B, \quad W_B = \frac{Q}{\mu(\mu - B)}. \quad (40)$$

A direct arrival also waits for direct applications already ahead of it and brokered applications that arrive during its wait. Thus

$$W_D = W_0 + \frac{B}{\mu} W_B + \frac{D}{\mu} W_D + \frac{B}{\mu} W_D, \quad W_D = \frac{Q}{(\mu - B)(\mu - Q)}. \quad (41)$$

Adding the common service time $1/\mu$ and subtracting gives

$$g(Q, B; \mu) = W_D - W_B = \frac{Q^2}{\mu(\mu - B)(\mu - Q)}. \quad (42)$$

This is the two-class exponential-service specialization of the nonpreemptive priority formula in Larson and Odoni, *Urban Operations Research* (1981), Section 4.9.1.

Comparative statics. For $0 < B < Q < \mu$,

$$g_Q = g \left(\frac{2}{Q} + \frac{1}{\mu - Q} \right) > 0, \quad g_B = \frac{g}{\mu - B} > 0. \quad (43)$$

Since the common recorded frontline time cancels from the gap,

$$\dot{\Delta} = g_Q \dot{Q} + g_B \dot{B} > 0 \quad \text{if } \dot{Q} > 0 \text{ and } \dot{B} > 0. \quad (44)$$

This proves Prediction 2 independently of the broker fee or any payment for priority. Citizen choice and broker entry may determine which flow changes occur; they cannot change this implication conditional on those flows.

Priority reallocates the same approval resource. In fact, the average approval time across both routes satisfies

$$\frac{B(1/\mu + W_B) + D(1/\mu + W_D)}{Q} = \frac{1}{\mu - Q}. \quad (45)$$

No additional approval capacity has been assigned to brokers. The result allows utilization strictly below one. Fixed approver headcounts alone are not sufficient: effective service capacity, service requirements, arrival assumptions, and the priority rule are also maintained. A changing priority rule or a productive broker service would constitute a different benchmark.

C.6. Derivation of the complementarity result

The productivity channel. From (17), delivering q_m applications requires $e_m = q_m(\omega_p/\xi)^\psi$ units of broker effort. The cost of this effort plus other per-client costs is $[c_0 + \bar{c}(\omega_p/\xi)^\psi]q_m$, giving the

stated marginal cost. Its equilibrium response is

$$\dot{c} = \psi(c - c_0) \frac{\dot{\omega}_p}{\omega_p}. \quad (46)$$

Thus a fall in total pre-submission interaction time lowers broker cost. The public work $b_B Q_B$ remains in (25); the cost saving concerns private effort, not an uncouneted supply of public labor.

To separate the three forces, substitute $P = c + r$ into broker demand and differentiate:

$$\dot{Q}_B = -\lambda \frac{V - c - r/2}{\tau^2} \dot{\tau} + \lambda \frac{c + r/2}{H^2} \dot{H} - \lambda h \dot{c}. \quad (47)$$

The first term is the participation response to common access, the second is the response to the assistance advantage, and the third is the broker supply response. The cost term is absent under substitution. All three terms use total equilibrium derivatives.

Applications and broker share. Suppose $x \geq 0$ and $y > 0$. From (33), direct volume rises if $k < y$, and broker share rises if $k > (V - P)(y - x)/V$. Requiring $k > 0$ also ensures that total applications rise. These are precisely the conditions in (18). Since $Q_B = S_B Q$, brokered volume then rises too.

The interval is nonempty for every $x \geq 0$ and $y > 0$: its lower bound is strictly below y because $(V - P)/V < 1$. Under the substitution bound, the constant-cost fee response is too small to cross the share-growth threshold. Equation (32) shows how a fall in c can raise k above that threshold. A sufficiently large fee decline can also violate $k < y$, however; lower broker costs alone do not guarantee increased direct volume.

Entry. Let $s = S_B$ and $\beta = P/(V - P)$. Accounting and (33) give

$$\varepsilon_{Q_B} = \frac{x - (1 - s)y + (\beta + 1 - s)k}{s}, \quad \varepsilon_M = \varepsilon_{Q_B} - \frac{\nu}{2}. \quad (48)$$

Hence entry rises exactly when

$$\varepsilon_{Q_B} > \frac{1}{2} \left[\frac{H}{\tau + H} x + \frac{\tau}{\tau + H} y \right]. \quad (49)$$

Equivalently, $k > k_E$, where

$$k_E = \frac{sv/2 - x + (1 - s)y}{\beta + 1 - s}. \quad (50)$$

The application and entry conditions can therefore be written together as

$$\max \left\{ 0, \frac{V - P}{V} (y - x), k_E \right\} < k < y. \quad (51)$$

This interval is nonempty if $k_E < y$, or equivalently $x + \beta y > sv/2$. That additional restriction makes explicit why entry is not automatic. It can hold within the substitution restriction: holding τ, H and x/y fixed, a sufficiently small positive baseline broker share makes the right-hand side small, while $\beta = (1 - s)H/\tau$ remains positive. This demonstrates a nonempty region of market-response conditions; the public delay technology must also satisfy the timing conditions below.

Recorded time and both public constraints. Because frontline work per application is fixed,

$$\dot{\Delta} = (\ell_D - \ell_B)\dot{\omega}_s, \quad \dot{T}_D = \ell_D\dot{\omega}_s + d_Q\dot{Q}, \quad \dot{T}_B = \ell_B\dot{\omega}_s + d_Q\dot{Q}. \quad (52)$$

The gap narrows if $\dot{\omega}_s < 0$. Direct applications become faster if

$$d_Q\dot{Q} < -\ell_D\dot{\omega}_s. \quad (53)$$

These conditions explicitly retain the extra approval congestion from $\dot{Q} > 0$. Brokered time is unchanged to first order when $d_Q\dot{Q} = -\ell_B\dot{\omega}_s$. At that equality, direct time still falls because $\ell_D > \ell_B$.

Prediction 3 follows by combining (51), $\dot{\omega}_p < 0$, $\dot{\omega}_s < 0$, and (53) on the interior, feasible equilibrium branch defined by (28). These are sufficient conditions for the joint response. They are not independent assignments of fees, flows, and times: all derivatives must come from that same equilibrium, with workload and approval congestion included as in (30). The analytical comparison identifies the cost force that can overturn crowd-out; it does not infer the unobserved preparation process from recorded times.

References

- A. Aneja and G. Xu. Strengthening state capacity: Civil service reform and public sector performance during the gilded age. *American Economic Review*, 114(8):2352–2387, 2024.
- N. Ashraf, O. Bandiera, E. Davenport, and S. S. Lee. Losing prosociality in the quest for talent? sorting, selection, and productivity in the delivery of public services. *American Economic Review*, 110(5):1355–94, 2020.
- P. Balán, A. Bergeron, G. Tourek, and J. L. Weigel. Local elites as state capacity: How city chiefs use local information to increase tax compliance in the democratic republic of the congo. *American Economic Review*, 112(3):762–97, March 2022. doi: 10.1257/aer.20201159. URL <https://www.aeaweb.org/articles?id=10.1257/aer.20201159>.
- A. Banerjee, E. Duflo, and G. Sharma. Long-term effects of the targeting the ultra poor program. *American Economic Review: Insights*, 3(4):471–486, 2021.
- P. Barnwal, A. Dar, A. Mahajan, M. Punjabi, and P. P. S. Sandhu. Managers and offices: Location-based determinants of bureaucratic effectiveness in india. Technical report, NBER Summer Institute 2026, Development Economics, Cambridge, MA, 2026.
- M. Bertrand, S. Djankov, R. Hanna, and S. Mullainathan. Obtaining a driver’s license in india: an experimental approach to studying corruption. *The Quarterly Journal of Economics*, 122(4):1639–1676, 2007.
- M. C. Best, J. Hjort, and D. Szakonyi. Individuals and organizations as sources of state effectiveness. *American Economic Review*, 113(8):2121–2167, 2023.
- E. P. Bettinger, B. T. Long, P. Oreopoulos, and L. Sanbonmatsu. The role of application assistance and information in college decisions: Results from the h&r block fafsa experiment. *The Quarterly Journal of Economics*, 127(3):1205–1242, 2012.
- S. Bhargava and D. Manoli. Psychological frictions and the incomplete take-up of social benefits: Evidence from an irs field experiment. *American Economic Review*, 105(11):3489–3529, 2015.
- A. T. Byrne, X. Dong, J. Handbury, E. James, K. Meckel, and A. C. Rovira. Welfare implications of increased retailer participation in SNAP. NBER Working Paper 34467, National Bureau of Economic Research, 2025.
- C. Chalendard, A. M. Fernandes, G. Raballand, and B. Rijkers. Corruption in customs. *The Quarterly Journal of Economics*, 138(1):575–636, 2023.
- E. Dal Bó, F. Finan, and M. A. Rossi. Strengthening state capabilities: The role of financial incentives in the call to public service. *The Quarterly Journal of Economics*, 128(3):1169–1218, 2013.
- A. Finkelstein and M. J. Notowidigdo. Take-up and targeting: Experimental evidence from snap. *The Quarterly Journal of Economics*, 134(3):1505–1556, 2019.
- A. Fredriksson. Bureaucracy intermediaries, corruption and red tape. *Journal of Development Economics*, 108: 256–273, 2014.
- A. J. Ganimian, K. Muralidharan, and C. R. Walters. Augmenting state capacity for child development: Experimental evidence from india. *Journal of Political Economy*, 132(5):1565–1602, 2024.
- S. Gupta. Perils of the paperwork: The impact of information and application assistance on social benefit take-up in india. *American Economic Journal: Economic Policy*, 18(3):342–371, 2026.
- K. Meckel. Is the cure worse than the disease? Unintended effects of payment reform in a quantity-based transfer program. *American Economic Review*, 110(6):1821–1865, 2020.

- K. Muralidharan, P. Niehaus, and S. Sukhtankar. General equilibrium effects of (improving) public employment programs: Experimental evidence from india. *Econometrica*, 91(4):1261–1295, 2023.
- M. Romero and A. Singh. The incidence of affirmative action: Evidence from quotas in private schools in india. *Research on Improving Systems of Education (RISE)*, 2022.