Privatized Provision of Public Transit*

Lucas Conwell †

First Version: October 2022
This Version: March 2026

Abstract

Privately-operated minibuses provide the majority of urban transit in developing-country cities yet are often viewed as a hindrance to development, not least because of often-long wait times passengers face for buses to fill up and depart. To quantify the efficiency of these seemingly-chaotic networks, I build the first model of privatized shared transit, which features increasing returns in the form of shorter waits on busier routes and a key role for bus supply as "insurance" against demand spikes. Minibus-sector market power inhibits realization of the former but facilitates internalization of the latter. I then estimate the model with newly-collected data on minibus arrivals and passenger queues in Cape Town as well as stated preferences for exogenously-varied commute attributes. Limited minibus market power in Cape Town allows demand and thereby wait times to reach an almost-efficient scale, suggesting that appropriately-regulated privatized minibuses can efficiently provide the public good of transit.

*I am grateful to my advisors Costas Arkolakis, Michael Peters, Mushfiq Mobarak, and Orazio Attanasio for their generosity with their time and ideas. I thank Gharad Bryan, Nina Harari, Allan Hsiao, Gabriel Kreindler, Oluchi Mbonu, Melanie Morten, Nick Tsivanidis, and Román David Zárate for their insightful comments and suggestions. I am also grateful to Treb Allen, Richard Blundell, Richard Carson, Alvaro Cox, Aureo de Paula, Mateus Dias, Fabian Eckert, Simon Fuchs, Jonas Hjort, Imran Rasul, Edouard Schaal, and Gabriel Ulyssea for their feedback which materially improved the paper. I am also grateful to the anonymous referees who provided useful and detailed comments on an earlier version of the manuscript. I thank Philip Krause, Aslove Mateyisi, and Mokgadi Mehlape from GoAscendal/GoMetro for their herculean efforts to obtain permission for and organize the data collection. Finally, I acknowledge generous financial support from the Stone Centre at UCL, the Yale Economic Growth Center, and the Ryoichi Sasakawa Young Leaders Fellowship Fund that made this project possible. IRB Exemption Determination obtained through Yale University, ID 2000032302.

†Department of Economics, University College London, London WC1H 0AX, UK. l.conwell@ucl.ac.uk

I. Introduction

In developing-country cities, the private sector often provides a substantial share of urban public goods—schools, health (Fabre and Straub (2023)), and waste collection (Banares-Sanchez and Wiskamp (2025))—that richer-city governments more often provide directly (Bryan et al. (2020)). This pattern appears most starkly in urban transport, where limited fiscal resources and state capacity often preclude investment in costly public transit infrastructure (Balboni et al. 2020). Into this gap, vast privately operated minibus networks have emerged: *trotros* in Accra, *cars rapides* in Dakar, *matatus* in Nairobi, and *micros* in Lima. Figure 1a underscores the extent to which the privatized transit share rises as city income falls, reaching 100% in some of the poorest cities (Tun and Hidalgo (2022)).

Yet the downsides of privatized transit reliance loom large. For one, minibuses clog already-congested roads and frustrate fellow road users with their erratic maneuvers. In the aggregate, workers in these cities, as evident in Figure 1b, waste significantly more time commuting (Kreindler (2022)) than their rich-country counterparts. In Cape Town—a sprawling, privatized-transit-dominated African city—average *one-way* minibus commutes reach 57 minutes, enough, given typical estimates of workers' responsiveness to commute times (Tsivanidis (2023) and Zarate (2024)), to meaningfully constrain their access to jobs. Wait time alone accounts for more than a third of this total. Commuters wait both in a queue for a minibus to arrive and on board, because the *fill-and-go* model near-universal to these *privatized shared transit* markets delays departure until buses fill.[1] Perhaps unsurprisingly, then, policymakers typically view minibuses as a hindrance to development — a view which economists, in their focus on publicly-provided alternatives, have rarely challenged. Nonetheless, minibuses connect far-flung suburbs at minimal public cost, including many low-demand routes with only slow and unreliable public transit options. This duality motivates my central question: can these seemingly chaotic privatized shared transit networks efficiently provide the public good of urban transit?

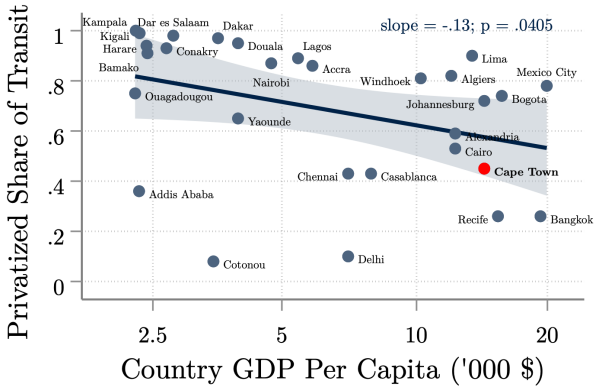
In this paper, I introduce what is, to my knowledge, the first model of privatized shared transit. Unlike standard urban models with exogenous commute costs, my framework endogenizes both the time and monetary costs of minibus travel. Higher commuter demand and bus supply reduce wait times, generating increasing returns, while equilibrium changes in wait times, road congestion, and fares interact with commuters' spatial sorting. Due to the scarcity of existing data on wait times and bus supply, I collected new data on passenger queues and bus arrivals in Cape Town. To estimate preferences over hard-to-observe minibus attributes, such as station security guards, I conducted discrete choice experiments that solve identification issues in the usual observational data. With the estimated model, I solve for the optimal commuter choices and minibus supply and find that, largely due to limited minibus-sector market power, Cape Town's decentralized minibus equilibrium lies close to the planner's optimum.

I begin with three facts about the wait times inherent to any form of shared transit, private or public. First, wait time accounts for 36% of total commute time on the average Cape Town minibus route, versus a worldwide median of 30% for public buses and trains. Second, passengers wait less on busier minibus routes, both in the

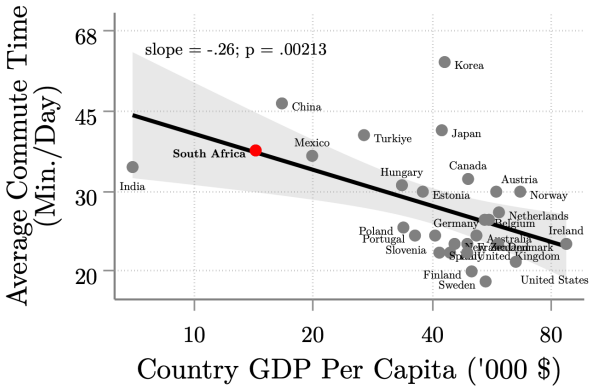
[1] Wait time accounts for a full one-third of the average transit commute worldwide (Nikolaidou et al. (2023)). A long line of work in urban planning discusses the costs of wait time for public (Abenoza et al. (2019), Ansari Esfeh et al. (2021), and De Vos et al. (2023)) as well as privatized transit (Cervero and Golub (2007) and Kerzhner (2022)).

FIGURE 1. URBAN TRANSPORTATION IN LOWER-INCOME COUNTRIES

(A) PRIVATIZED SHARE OF TRANSIT



(B) COMMUTE TIME



Notes: Panel (A) displays a scatterplot at the city level of the share of shared transit (rail, buses, and other multi-person vehicles) accounted for by private providers from Tun and Hidalgo (2022), versus national GDP per capita; Panel (B) displays a scatterplot of average countrywide commute time from the OECD (2016) versus national 2019 PPP GDP per capita in current international dollars from the International Comparison Program (2020).

queue until a bus arrives and on board before departure under the near-universal *fill-and-go* model. Third, demand shocks generate queues, which higher bus supply "insures" against. These relationships generalize to any (small-vehicle) transit; the nature of provision, private versus public, only determines the degree to which routes realize these increasing returns.

I then build a new model of privatized shared transit that, unlike standard frameworks with fixed "iceberg" commute costs, captures the equilibrium responses of wait times and fares. On each origin-destination route, a single minibus association chooses bus supply and Nash-bargains fares with the city government, generating a limited form of market power. A queueing model, tailored to minibus operations and solved out of steady-state, determines wait times: passengers first queue until a minibus arrives, then wait on the bus until it fills up and departs. Travel times, in turn, rise with road congestion.[2] Commuters with heterogeneous incomes choose a transport mode as well as home and work locations based on factors such as commute times and safety.

My model highlights three potential sources of inefficiency. First, minibuses' strict capacity constraint creates a technological link between demand and waits: twice as many passengers require twice-as-frequent departures. I term this source of increasing returns the *Market Mohring Effect,* in line with the economies of scale from which all forms of shared transit benefit (Mohring (1972)). Second, minibuses' low capacity also introduces the potential for queues of passengers to build up, which higher bus supply helps dampen in a novel *Bus Supply Insurance Effect.* Third, minibuses and cars generate standard road congestion externalities. In this setting, market power is a double-edged sword. On the one hand, greater bargaining power allows associations to charge higher fares, which suppresses demand and thereby inhibits Market-Mohring increasing returns. On the other hand, higher fares boost the marginal return to better service quality in the form of

[2] I abstract away from stops en route, which Cape Town's highway-reliant road network largely prevents.

additional Bus Supply "Insurance."

I collected two types of primary data in Cape Town to measure wait times and identify preferences over typically unobserved commute attributes. First, enumerators recorded passenger queues and bus arrivals on a random sample of 44 minibus routes, from which I calculate wait times. Second, I conducted discrete choice experiments for commute mode choice, a familiar decision well-suited to stated-preference methods. In my survey, 526 respondents chose among hypothetical minibus commute options with exogenously-varied travel times, costs, and quality improvements, e.g. station security. Preferences across different modes come from a separate city-run stated preference survey.

I then estimate the model's structural parameters in four steps. First, I estimate passenger queueing efficiency from the relationship between queues and bus loading times. To address unobserved loading interruptions, e.g. due to weather, I instrument queue prevalence with past commuter work start times, likely uncorrelated with contemporaneous interruptions. I also estimate bus arrival efficiency from the relationship between bus supply and inter-arrival gaps, instrumenting supply with route distance. Second, the stated preference data circumvents the usual endogeneity issues inherent to travel survey data and thereby identifies mode-specific utility costs and the value of time. Third, I externally calibrate secondary parameters and the model geography. Fourth, I internally calibrate the remaining supply parameters and location fundamentals.

Finally, I use the estimated model to evaluate the efficiency of Cape Town's decentralized minibus equilibrium. Nonzero association bargaining power vis-a-vis city government keeps fares high enough to suppress demand below the optimal scale, so the planner would raise minibus demand by 15% on the median route to exploit the Market Mohring Effect and reduce waits. At the same time, because bargaining power remains below one, associations do not fully internalize the Bus Supply Insurance Effect, so the planner would increase minibus supply by 46% on the median route. In practice, buses arrive frequently enough that the Market Mohring channel dominates. Limited association bargaining power, in turn, allows decentralized equilibrium demand to approach its optimal scale, so the planner achieves only marginal welfare gains. I then decompose minibus demand and show that commuters face a trade-off between short commute times and crime risk, driving safety, and comfort. Thus, the case of Cape Town teaches a more general lesson: privatized providers with appropriately-limited market power can indeed efficiently provide the public good of urban transit.

Related Literature

A large literature studies the effects of urban *public* transport infrastructure in developing countries. These papers typically exploit new government-run bus rapid transit or metro lines as natural experiments to estimate key parameters of quantitative spatial models (Ahlfeldt et al. (2015)) and trace effects on worker sorting (Tsivanidis (2023)), gentrification (Balboni et al. (2020)), labor informality (Zarate (2024)), or crime (Khanna et al. (2024)). Kreindler et al. (2023) go a step further and characterize optimal public transit networks. These models successfully predict commuters' location choices as well as spillovers and aggregate effects but generally treat commute costs as exogenous "iceberg" objects. This assumption limits their applicability to privatized shared transit, where wait times, travel times, and fares respond endogenously to passenger and provider choices. I contribute the first model of this privatized transit sector and thereby quantify these

equilibrium responses as well as the scope for efficiency gains.

More recent work uses natural experiments rather than structural estimation to address related but distinct questions. Björkegren et al. (2025) study how privatized shared transit responds to Lagos's public-bus rollout and derive model-based sufficient statistics for welfare. Mbonu and Eaglin (2024) estimate shorter-run minibus supply responses to demand shocks and bus breakdowns and emphasize the role of market structure.

More generally, I relate to a broader literature which applies structural models to study developing-country cities (Bryan and Morten (2019), Kreindler (2022), Hsiao (2023), Gechter and Tsivanidis (2023), Vitali (2024), Harari and Wong (2024), and Hsiao (2024)) and to the public-private partnership literature reviewed by Fabre and Straub (2023). The latter body of papers, often focused on education or health, study contract design and optimal ownership structures (Besley and Ghatak (2001)). I extend classic questions about the efficiency of private provision (Norman (2004)) to a new domain: the public good of shared transit.

I structure the remainder of the paper as follows. In Section II, I describe my data and the institutional context. In Section III, I discuss a series of facts to rationalize my focus on wait time. I then lay out my theory in Section IV, followed by the estimation procedure in Section V. After I validate my model's fit in Section VI, I compare the decentralized minibus equilibrium to the optimum in Section VII. Finally, I conclude in Section VIII.

II. Minibus Data Collection and Context

In this section, I discuss the collection of my primary data and then provide an overview of the minibus market in Cape Town. Online Appendix B provides additional details regarding all datasets used in the paper.

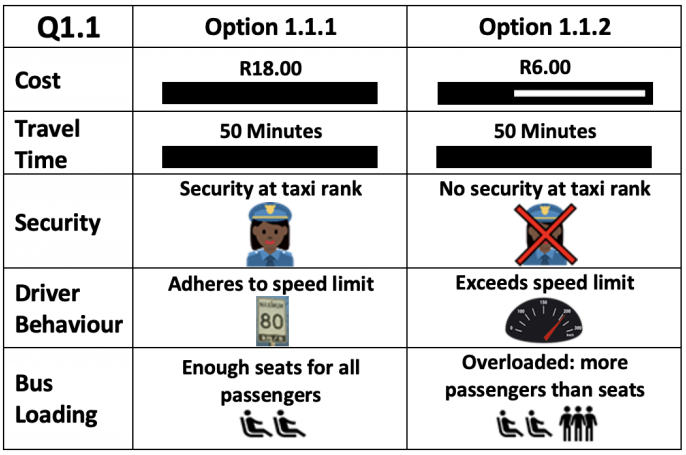
Newly-Collected and Household Survey Data

Because no suitable data exists on minibus passenger wait times and quality, I conducted a custom two-part data collection effort in Cape Town. First, I quantify the supply side of the market via minibus *station counts,* which tracked the process by which these buses load passengers in route-specific lanes at origin stations. For a sample of routes, enumerators recorded bus arrival and departure times, the number of passengers on board each bus, and, at five-minute intervals, the length of the queue to board a given route, all over the course of the 6-10am morning peak commute. With these observations in hand, I impute passengers' wait times, as detailed in Online Appendix B.1.3. I pair these observations of the loading process with data on fares, passenger numbers, and travel times collected on board randomly-sampled minibus trips on these same routes.

These counts covered a two-stage cluster sample of origin stations and then routes from a roster developed by the City of Cape Town.[3] Within each stage, I stratified by a proxy for bus supply, namely the number of

[3] In order to update the municipal Transport and Urban Development Authority's record of on-the-ground minibus route paths and operations, enumerators armed with a smartphone app made close to 30,000 minibus trips on approximately 800 established minibus routes. The results showed the official, city-designated routes to be outdated. For example, 250 of the official city routes no longer operated (Coetzee et al. (2018)). This listing is, according to stakeholders, as comprehensive and accurate as any available and includes the number of minibus trips mapped by route in this previous data collection effort. Stakeholder discussions revealed that the number of trips mapped is an indicator of the number of buses on a route.

FIGURE 2. MINIBUS STATED PREFERENCE SURVEY: CHOICE SET



Notes: This figure shows an example of a choice set from my stated preference survey, consisting of two hypothetical minibus commute alternatives, from which respondents indicated their preferred option. The rows list the attributes associated with each option, which vary exogenously across choice sets and respondents. Note that "taxi rank" is the South African term for a minibus station.

trips mapped in the creation of the aforementioned roster, to reduce the likelihood of sampling no longer operational routes. The sampling frame excluded stations and routes where local minibus associations denied permission as well as those without morning peak operations; when such impediments came to light only ex-post, I re-sampled new stations or routes.[4] My final two-stage cluster sample comprises 44 routes. Despite the aforementioned practical constraints, the routes I sample within each stratum, as I lay out in Online Appendix Table B.1, broadly mirror the corresponding subset of the 429 unique routes in Cape Town in terms of bus traffic, length, and fares paid.

Second, I quantified commuter demand via a stated preference survey composed of a series of discrete choice experiments designed to estimate commuters' values of time and quality as well as price sensitivity. In my survey, I asked respondents to consider a hypothetical work commute trip and then choose a preferred option in a series of *choice sets* composed of two minibus alternatives, as in Figure 2. Options varied in five attributes: cost, travel time, and three binary hypothetical quality improvements. The latter, chosen based on commuter concerns as summarized in Online Appendix Figure A.1, included security guards at the (publicly-owned) minibus stations, driver adherence to speed limits, and whether the minibus takes on more passengers than seats. I restricted attributes to take on realistic values and chose the bundles of attribute values to maximize parameter precision via a *d-efficiency* algorithm now standard in the literature (Rose and Bliemer (2009)). Respondents then completed one of two randomized questionnaires of five choice sets each.

Sampling for the survey occurred in-person in order to economize on resources. Security considerations did not permit a random sampling of minibus stations or other locations. I instead chose three locations, namely the Bellville/Middestad Mall intermodal transport interchange and the Khayelitsha Site C and Somerset West

[4]In the field, 14 sampled routes unexpectedly turned out not to operate at all during the morning peak. Field supervisors then randomly selected replacement routes at the same station on the spot, with the exception of one station, which turned out to have only 2 total routes in operation upon commencement of the day's data collection, such that no replacement routes could be sampled.

Shoprite minibus stations, to capture a broad cross-section of commuters. At these three sites, enumerators randomly approached a total of $N = 526$ respondents on a series of weekdays. While enumerators did not explicitly track refusals, anecdotal reports suggested a high willingness to participate, and all 526 employed respondents who began the survey also completed all questions.[5]

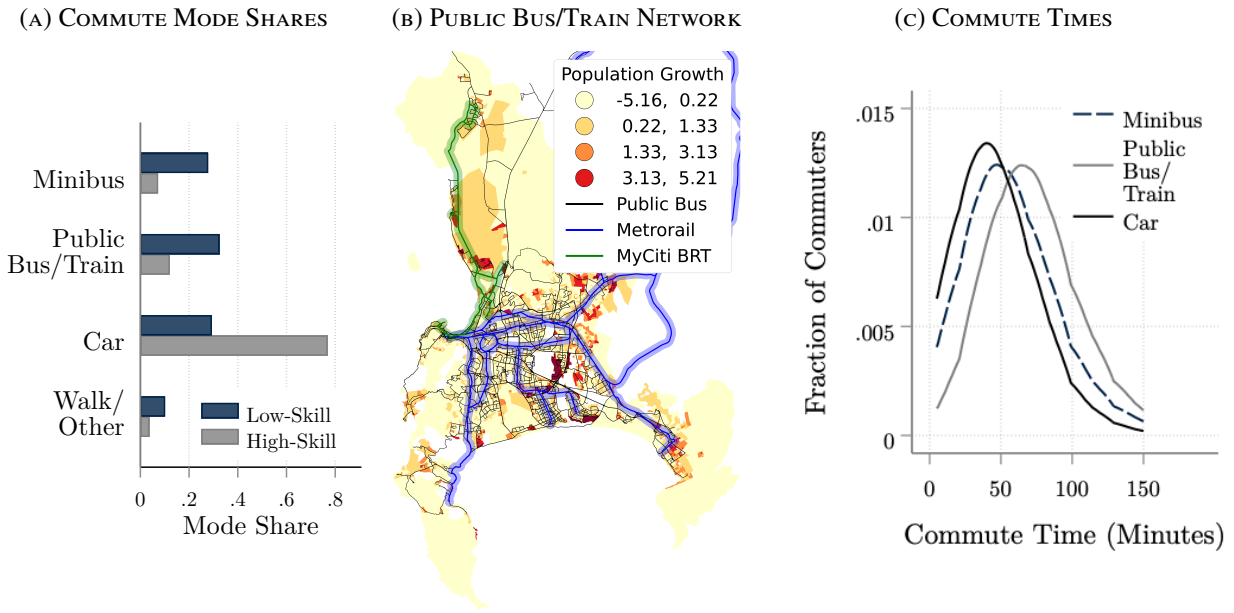
While my sample mirrors the entire Cape Town commuter population along standard demographic dimensions such as age, gender, education, and income, the distribution of respondents' actual typical commute modes differs from that of the population. In particular, my survey over-sampled commuters who report typically using minibuses, as detailed in Online Appendix B.2.6. This sample selection might be most likely to bias estimates of relative preferences across these different modes of transport; my survey, however, features only minibus options and so does not contribute to the identification of different modes' relative utility. Furthermore, I confirm in Online Appendix B.2.6 that minibus users citywide have similar observable characteristics to public bus or train commuters and differ from car commuters primarily in terms of education and income, both of which my model will capture. Nonetheless, I later test the robustness of my demand estimates to weighting by realized commute mode. To more directly investigate whether the interview locations themselves might have induced selection, I compare respondents at different interview locations, also in Online Appendix B.2.6, and find little variation in observables. To probe for selection on *un*observables, I later re-estimate the model separately for each type of location and also analyze heterogeneity in demand elasticities and values of security across these different subsamples.

The literature has identified two key challenges specific to stated preference approaches: comprehension and *hypothetical bias,* or distortions related to the fact that respondents do not need to satisfy an actual time or monetary budget constraint. Commute mode choice arguably corresponds closely to the kind of "choice[s] among a small number of realistic, familiar, and fully described products" most likely to facilitate respondent comprehension (Ben-Akiva et al. (2019, p. 13–14)). Nonetheless, to test the cognitive load on respondents, I conducted a pilot survey, after which I reduced the number of choice sets and alternatives per set. In line with best practices in developing-country contexts (Mangham et al. (2009)), I chose relevant attributes based on past surveys, included pictograms, and ensured that trained enumerators guided respondents question-by-question. A common form of hypothetical bias, in turn, arises when respondents distort their answers to direct questions regarding willingness to pay because they hope to influence the implementation of particular policies (Whittington (2010)). My discrete choice experiments, which involve alternatives which vary in multiple dimensions, would require an unusual degree of sophistication to "game" in this fashion. Nonetheless, as a test for such bias, I follow state-of-the-art guidance (Whittington (2010) and Ben-Akiva et al. (2019)) and collect respondents' *revealed* preferences, i.e. actual commute modes, for later comparison to their stated preferences.

Finally, I pair my stated preference survey with household survey data collected by the City of Cape Town. The 2013 Cape Town Household Travel Survey provides home and workplace locations as well as incomes for a representative citywide sample of $N = 22,332$ households. A smaller subsample completed another

[5]The in-person, in-public survey mode, chosen to ensure that respondents carefully completed and comprehended the survey, precluded reliable calculation of response rates and, naturally, the collection of data on non-respondents who refused to speak to enumerators.

FIGURE 3. MODES OF TRANSPORTATION IN CAPE TOWN



Notes: Panel (A) displays the shares of low- (non-university) and high-skill commuters in Cape Town who use each mode, as measured in the 2013 Cape Town Household Travel Survey. I exclude residents who work in their home transport analysis zone, and "minibus" includes any who use minibuses during a typical commute. Panel (B) displays the networks of publicly-provided transit modes with dedicated infrastructure in Cape Town, namely MyCiti bus rapid transit, Metrorail commuter trains, and the standard Golden Arrow public bus. Shading indicates population growth from 1996–2011 at the small area layer level. Panel (C) displays the distributions of commute time across commuters by primary mode of transport used.

commute stated preference survey, which, unlike mine, involved hypothetical choices between different modes of transport: car, public bus or train, or minibus. These options varied in commute time and cost, but not, as in my survey, quality-related attributes.[6]

Minibuses in Cape Town

Cape Town provides an insightful setting in which to study privatized shared transit. Though its apartheid history induced particularly extreme forms of spatial segregation, Cape Town's sprawling urban geography (Anderson et al. (2023) and OECD/SWAC (2023)) typifies the unusually fragmented cities of sub-Saharan Africa (Angel et al. (2016)), which public metro or bus rapid transit networks inherently struggle to serve efficiently.[7] Cape Town's minibuses, in turn, ply defined routes managed by state-sanctioned monopolies. While these arrangements belie the "informal" moniker often applied such networks, a similar mix of sophisticated collective organization, market power, and government regulation characterize minibus markets across the African continent (Kerzhner (2022, 2023)).

[6]This city-run stated preference survey, like my own, relied on low-dimensional, familiar choice sets, pictograms, and in-person interviews to boost respondent comprehension and thereby achieved an overall response rate around 90%.

[7]Indeed, Africa is the only continent whose cities included in the Atlas of Urban Expansion are growing *more* fragmented, and when the Atlas was published, their average fragmentation level was already high compared to European, Japanese, and Latin American cities where publicly-provided transit has been successful (Angel et al. (2016)). On two new indices of sprawl and elongation (OECD/SWAC (2023)), Cape Town falls close to the African average (Anderson et al. (2023)).

FIGURE 4. MINIBUS LOADING PROCESS

(A) PASSENGER QUEUE



(B) BUSES LOADING, ONE-AT-A-TIME



Notes: Panel (A) displays an example of a passenger queue to board a specific origin-destination route at its designated loading lane within the origin station, here the Cape Town CBD station. Panel (B) displays the Mfuleni minibus station with one bus per route loading in the corresponding designated loading lane. Images by author, June 2022.

Privatized shared transit, almost exclusively in the form of late-model Toyota Quantum/HiAce 15-passenger minibuses, provides a lifeline to lower-income commuters in Cape Town. Figure 3a displays the shares of commuters in Cape Town who use each mode of transport. A full 28% of non-university-educated, henceforth "low-skill," and 7% of high-skill workers commute via minibus. Another third of low-skill commuters use publicly-provided bus or train alternatives, which include Golden Arrow buses running in mixed traffic, aging Metrorail commuter rail lines, and a single higher-speed bus rapid transit (BRT) line known as MyCiti. Finally, a non-trivial fraction of low-skill as well as the overwhelming majority of high-skill commuters instead drive to work. Almost all neighborhoods have access to some form of public bus or train, as evident in the combined networks overlaid on population growth in Figure 3b. However, since Cape Town has only built a single BRT line, the public "outside" option in most neighborhoods consists of typically slow, infrequent non-BRT buses and trains. Indeed, the distributions of commute times by mode in Figure 3c demonstrate minibuses' advantage in terms of commute times.[8]

The minibus market consists of a large number of small private firms. A well-developed ecosystem of commercial banks, a minibus-specific financier, and a nationwide *Taxi Recapitalization Program* provide loans for investment in new minibuses (Commission (2020), Schalekamp and McLachlan (2016), and Schalekamp and Klopp (2018)).[9] Firms own, on average, less than two buses each (Woolf and Joubert (2013)) and typically hire drivers distinct from owners. Around three-quarters of minibus drivers in South Africa earn a fixed wage, although, at some individual stations in Cape Town, majorities reportedly receive a commission based on fare revenue (Antrobus and Kerr (2019)).

Minibus firms, as in many sub-Saharan African cities (Kerzhner (2022)), collude via minibus *associations*.

[8]Nonetheless, the same household survey data indicates that public bus/train commuters pay 57% less per trip, on average, than minibus commuters.

[9]Indeed, thanks in part to the Recapitalization Program as well as the high returns in the sector (Kerr (2018)), one-third of minibuses nationwide have reportedly been upgraded since the program's inception (Schalekamp and Klopp (2018)).

Cape Town city government typically grants a single association per origin-destination pair, or *route,* the exclusive rights to set fares, subject to government approval (City of Cape Town (2014, p. 65)), and regulate firm entry (Kerr (2018)). Individual firms then pay an entry fee to such an association to operate on a single route. In contrast to public-private partnerships in developed-country contexts (Besley and Ghatak (2001)), the government serves primarily as a regulator and does not take a stake in minibus firms, own any vehicles, or provide subsidies beyond the construction of stations.

Daily operations, in turn, mimic practices common across developing-world cities, whether Djibouti, Nairobi, or Lilongwe (Kerzhner (2022)). Minibuses in Cape Town depart at random, unscheduled times from large dedicated stations.[10] Stations feature clearly marked loading lanes for each route, where passengers wait in sometimes-sizable queues to board minibuses, as pictured in Figure 4a. Figure 4b depicts the typical case where one minibus per route loads passengers from the front of the queue. In my data, 96% of these minibuses follow the *fill-and-go* practice near-universal among privatized shared transit providers worldwide (Cervero and Golub (2007)) and depart with a full load of at least 15 passengers.[11] After one bus departs, limited bus supply and congestion in and around stations often mean that a gap of several minutes passes before another bus arrives to load passengers. During the rush hours I study, minibuses typically provide a direct origin-destination service, often on highways with few logical intermediate stops.[12]

III. Facts About Waiting for Minibuses

Why might wait time, as opposed to the plethora of other frictions which plague privatized shared transit markets across the developing world, offer particular scope for gains? Shared transit of any form, from Uber Pool all the way to the Tokyo Metro, necessarily requires its users to wait, and Cape Town's minibuses are no exception.

Fact 1. *Wait time accounts for 36% of total commute time on the average Cape Town minibus route – and particularly plagues suburban routes.*

Across routes in my station count data, passenger wait times—*queueing time* plus the *loading time* until the bus fills up and departs—account for over a third of the 43-minute average total minibus commute. Wait times consume a particularly large fraction of suburb-to-suburb minibus commutes, as evident in Figure 5a. Indeed, once finally underway, minibuses frequently take highways and make few intermediate stops.[13] The wait time friction certainly generalizes to all forms of transit, public or private. Figure 5b displays the distribution of wait times for *public* buses and trains, as a fraction of total commute times, across a global

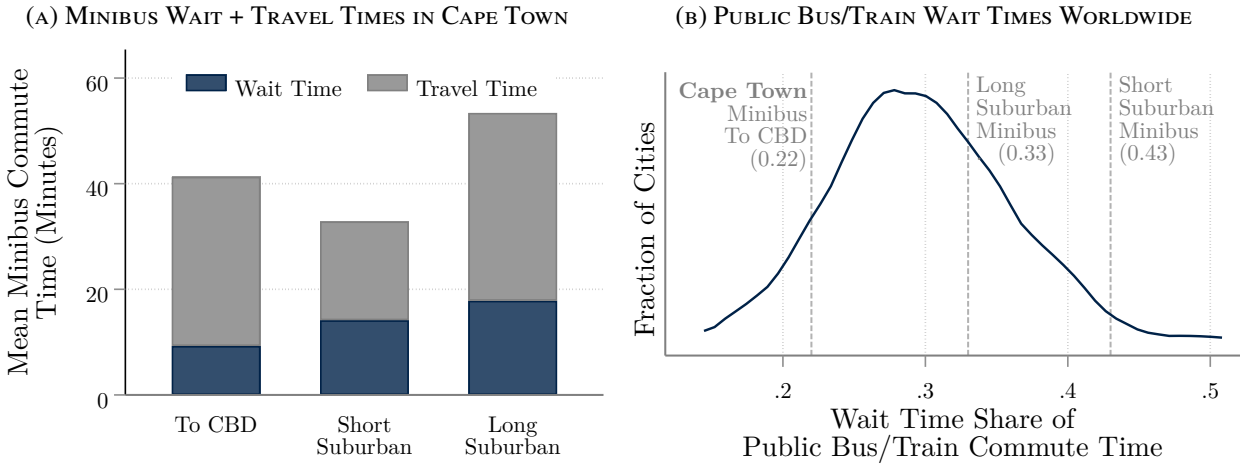
[10] In Online Appendix Figures A.2a-A.2b, I investigate the minutes of the hour at which minibuses on two representative routes depart and find no systematic patterns that would evidence even an informal, word-of-mouth schedule.

[11] Online Appendix Figures A.3a-A.3b display the distributions of queueing time and the number of buses loading simultaneously in my station count data; during 70% of the peak commute hours I observe, exactly one minibus is loading passengers. Though the law allows for several discrete bus sizes, the 15-passenger-plus-driver variant accounts for 94% of licensed minibuses (Jobanputra (2018, p. 290)).

[12] In fact, in my on-board tracking data, 80% of passengers board at the origin and 67% leave the bus at one of the last three stops, close to the destination.

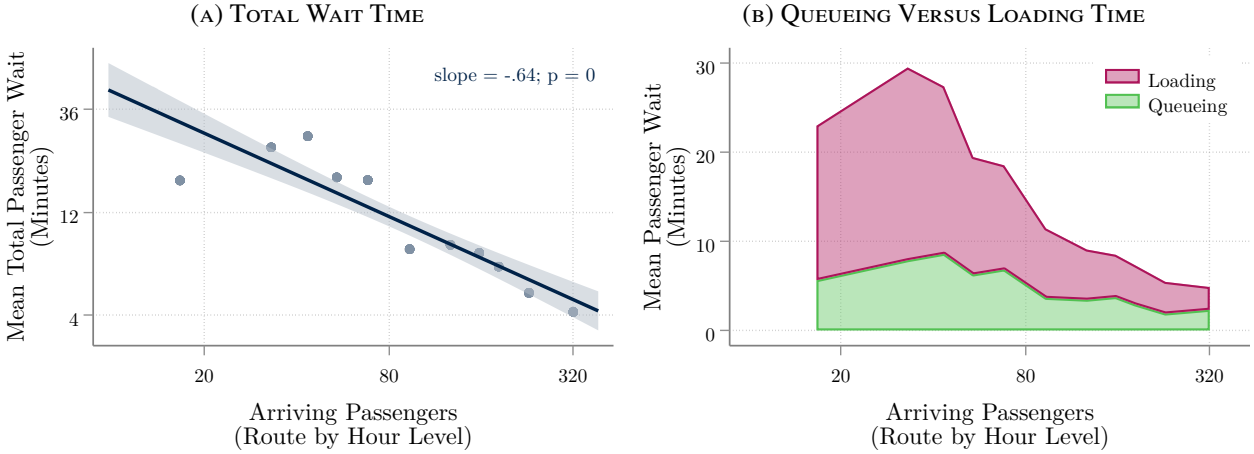
[13] In my sample, minibuses achieve an end-to-end average speed of 36 kilometers per hour. In comparison, buses in New York City recently averaged 13km/h.

FIGURE 5. WAIT TIMES



Notes: Panel (A) displays a stacked bar graph of the mean wait time (mean passenger queueing time plus mean loading time for the bus to fill up) and travel time in my station count data on routes of a given type. *To CBD* indicates routes which terminate in the transport analysis zone containing the CBD, while *Short* and *Long Suburban* are those that do not terminate in the CBD, with average distance driven below and above the median, 13.5km, across suburban routes. Panel (B) displays a histogram of average public bus/train wait times across 50 cities worldwide as calculated in the Moovit Global Public Transport Report 2024 (Moovit (2024)).

FIGURE 6. MARKET MOHRING EFFECT



Notes: Panel (A) displays binned scatterplot and best fit line for the log-scale relationship between the number of arriving passengers per hour for a given minibus route and hour and mean total passenger wait time, which includes queueing time and the mean minibus loading time from the station count data. Panel (B) separately plots (levels of) mean queueing and loading times by quantiles of passenger arrivals per route and hour.

sample of cities (Moovit (2024)), and the different types of minibus routes in Cape Town, as indicated, span the range of global wait times.

Mohring (1972) famously observed that all of this waiting introduces a particular form of increasing returns to any kind of shared transit technology, whether private or public, large or minibus. In Cape Town,

Fact 2. *Passengers wait less on busier minibus routes, due to both lower queueing and lower loading times.*

Figure 6a plots the relationship between, on the horizontal axis, the rate of newly-arriving passengers for a given route and hour and, on the vertical axis, mean total wait times, and the latter clearly fall with passenger arrivals. Indeed, minibuses' typically-binding capacity constraint bakes what I term a *Market Mohring Effect* into the technology itself: routes with twice as many passengers require twice as many buses, yielding twice the frequency and half the average wait time.[14] In other words, associations must add buses to accommodate the additional passengers, which decreases queueing time, and these buses then fill up in a shorter loading time, as both evident in Figure 6b. Despite stark differences in the underlying market structure, the Market Mohring elasticity of waits to passengers in Cape Town's minibus network, at -0.64, comes remarkably close to the -0.67 estimated for developed-world public buses and trains by Parry and Small (2009).[15] More generally, the nature of public or private provision determines not so much the slope of the Market Mohring curve in Figure 6a but rather, via the fares charged, the point on this curve reached in equilibrium.

In the case of small transit vehicles, with their typically-binding capacity constraints, higher demand also comes with a cost: the potential for queues. Indeed, over the course of the day, the total number of seats on minibuses—if they depart full—must equal the number of passengers, but spikes in passenger arrivals may not all fit on the next minibus. In consequence,

Fact 3. *Passenger arrivals lead to the build-up of queues, but additional minibus supply acts as "insurance" and dampens queues.*

Figure 7a plots passenger arrivals on the horizontal axis and the resulting number of passengers queueing after five minutes on the vertical axis. Sudden demand spikes evidently increase queue length in a convex fashion. Figure 7b, in turn, shows that total passenger wait times indeed fall with the number of buses operating on a route. Intuitively, the greater bus supply, the more frequently buses arrive, which decreases the likelihood that sudden demand shocks translate into long queues. I henceforth refer to fact that additional buses help weather demand shocks and prevent the build-up of queues as the *Bus Supply Insurance Effect*. Here again, fundamentals, i.e. bus size and urban geography, determine the *relationship* between bus supply and queues, while the degree of public ownership or the extent of private market power determine the *quantity* of buses supplied in equilibrium.

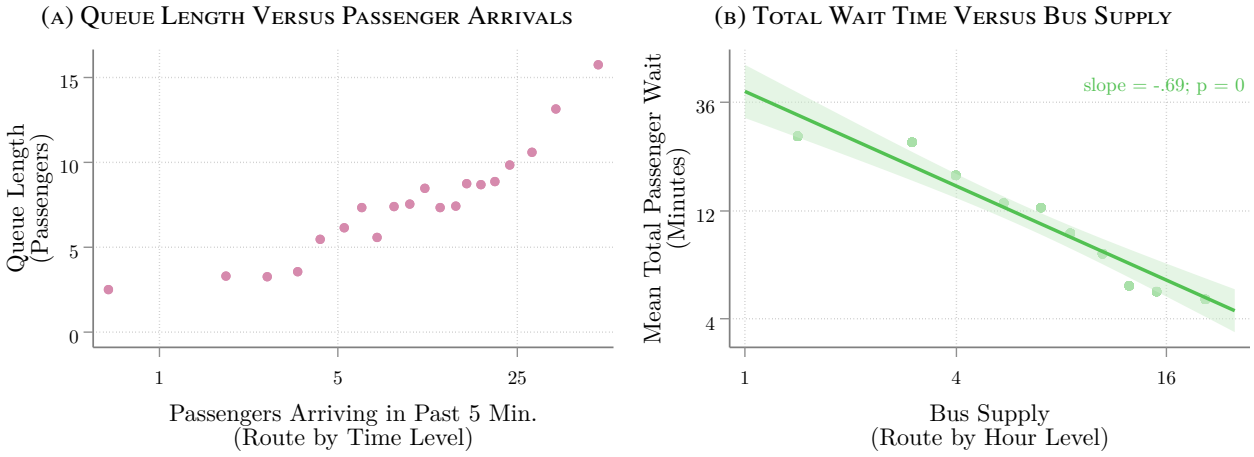
IV. A Theory of Privatized Shared Transit

In this section, I build a model of privatized shared transit. Passenger wait times come from a canonical M/M/1 queueing model (Hillier and Lieberman (2024)) adapted to minibus loading. On the supply side,

[14] In Online Appendix Table B.2, I compare commuter characteristics across minibus routes with above- versus below-median total daily passenger arrivals; commuters on high-demand routes are broadly similar to those on lower-demand routes in terms of most observable characteristics, including gender, education, age, home and work distance from the CBD, and time leaving home to go to work, but have somewhat lower median incomes.

[15] The comparable elasticity in Parry and Small (2009) equals -0.67, which I obtain by multiplying their elasticity of wait times with respect to vehicle frequency, at headways below 15 minutes, of -1, with the calibrated fraction of additional public transit demand accommodated through additional vehicles, 0.67.

FIGURE 7. BUS SUPPLY "INSURANCE" EFFECT



Notes: Panel (A) displays a binned scatterplot of route queue length in number of passengers versus the number of passengers arriving within the previous 5 minutes, across routes and every-five-minute snapshots on a log scale. Panel (B) displays the relationship between the number of buses operating on a given route during a given hour and mean total passenger wait time, which includes queueing time and the mean minibus loading time. All data comes from the station counts.

route-level minibus associations Nash-bargain fares with the city government and then choose bus supply. On the demand side, commuters of heterogeneous skill choose a home location, work location, and transport mode in a parsimonious quantitative spatial model in the tradition of Ahlfeldt et al. (2015). I first describe the environment, then characterize wait times and each agent's problem, define equilibrium, and then derive its efficiency properties.

Environment

I consider a city made up of a finite number of locations indexed by $i, j \in \{1, \ldots, I\} \equiv \mathscr{L}$ and two classes of agents: minibus associations and commuters. Each origin-destination pair (i, j), or what I henceforth term a "route," constitutes a separate minibus market, in which one association acts as a constrained monopolist. On each route, the association first Nash-bargains the fare with bargaining power β vis-a-vis the city government and then freely chooses the supply b_{ij} of minibuses. Minibuses have a fixed, common capacity $\overline{\eta}$. Fixed masses N^g of commuters of each skill g are born per unit time. Commuters choose a home location, indexed by i and with fixed amenity θ_i^g, a work location, indexed by j and with fixed wage ω_j^g, and one mode m, either minibus, M, public bus/train, P, or car, A. Locations are connected by a road network: the nodes consist of all locations $\mathscr{L}$, and links connect each location i and some fixed set of neighbors $S(i)$. Congestion in the road network determines minibus and car travel times, while public bus/train wait times, travel times, and fares are exogenous. Time is continuous, and commuters discount the future at rate r.

Waiting for Minibuses

I build on a canonical M/M/1 queue to model two components of minibus passenger wait time: (i) *queueing time* before boarding and (ii) *loading time* on board while the bus waits to fill and depart. For tractability, both passengers and bus arrivals and departures from the queue are stochastic. Passengers on route ij enter a route-specific first-in-first-out queue at Poisson rate λ_{ij}, pinned down by commuter demand. Whenever a bus is present in the loading area, passengers at the front of the queue board at Poisson rate μ. This *queueing efficiency* reflects physical constraints, e.g. the width of station platforms, as well as behavioral and cultural factors and equals the inverse of the average time each passenger takes to board.

A queueing model for minibuses must account for a key physical reality: passengers can board only when a bus is physically present at the stand. Following Avi-Itzhak and Naor (1963), I therefore augment the standard M/M/1 queue with "service interruptions," i.e. intervals with no bus available for boarding. Specifically, after one bus departs, the next arrives at Poisson rate $\lambda_{ij}^B \equiv \mathrm{g}\left(b_{ij}, L_{ij}, T_{ij}; \rho\right)$. The function g maps bus operations into arrival frequency: for given bus supply b_{ij} and *loading time* L_{ij} spent taking on passengers, longer route travel time T_{ij} slows bus turnover and thereby slows arrivals back to the origin. Moreover, station-area congestion implies that, even on high-supply routes, a new bus might not instantaneously replace a departing vehicle. The *arrival efficiency* parameter $\rho < \infty$ governs how quickly higher supply translates into faster replacement at the stand. Higher ρ corresponds to milder station congestion and thus a stronger pass-through from b_{ij} to λ_{ij}^B. Once a bus arrives, boarding resumes.

To capture "fill-and-go" operations while preserving stationarity, I assume buses depart at a Poisson rate μ_{ij}^B which adjusts so that the average departing load of passengers equals capacity $\overline{\eta}$.[16]. A useful benchmark is a continuously nonempty queue: if passengers are always available to board, boarding occurs at rate μ, so departures at rate $\mu/\overline{\eta}$ deliver average passenger load $\overline{\eta}$. With Poisson passenger arrivals, however, the queue is sometimes empty, so effective filling occurs only when both (i) a bus is present and (ii) at least one passenger is waiting. Let n_{ij} denote the number of passengers in the queue, and let $p_{ij} \equiv \Pr\left(n_{ij} > 0 | \text{bus present}\right)$, or the probability that passenger queue length exceeds zero, conditional on a bus being present in the loading area. Then, effective boarding flow while a bus is present equals μp_{ij}, so the bus departure rate consistent with full departures satisfies $\mu_{ij}^B \equiv \mu p_{ij}/\overline{\eta}$. Given the usual Poisson properties, a bus's expected loading time spent filling up at the origin station is

$$L_{ij} \equiv \frac{1}{\mu_{ij}^B} = \frac{\overline{\eta}}{\mu p_{ij}}. \tag{1}$$

Intuitively, higher p_{ij} raises effective filling speed, increases μ_{ij}^B, and lowers loading time L_{ij}.

Given passenger demand λ_{ij} and the bus arrival rate λ_{ij}^B, this queueing model implies route-level expected wait times. Following Brancaccio et al. (2024), I simulate queue dynamics out of steady state, as I detail in Appendix A.1. For each route ij, the simulation returns the probability of a nonzero queue conditional on bus presence, $p_{ij} \equiv p\left(\lambda_{ij}, \lambda_{ij}^B\right)$, and thus pins down both wait time components: (i) *queueing time*

[16] In Online Appendix E.4, I consider the scope for the social planner to attain additional gains from allowing buses to depart earlier, i.e. with lower $\overline{\eta}$.

$Q_{ij} \equiv Q\left(\lambda_{ij}, \lambda_{ij}^B\right)$, defined as the expected time from passenger arrival to the completion of boarding; and (ii) loading time $L_{ij} \equiv L\left(\lambda_{ij}, \lambda_{ij}^B\right)$, as in Equation (1).

Steady-State and Special Cases

For intuition, I now solve the queueing model in steady state. In this context, steady state means that inflows and outflows balance, on average, so that expected queueing and loading times remain constant over time. Two steady-state identities, derived in Appendix A.2, summarize the model. First, the *conditional* probability of a nonzero queue, given a bus present in the station, satisfies

$$p_{ij} = \frac{\frac{\lambda_{ij}}{\mu}}{1 - \frac{\lambda_{ij}}{\overline{\eta}\lambda_{ij}^B}}.$$

The numerator corresponds to the joint probability of a bus being present and a nonzero queue. By standard steady-state queueing properties, this share of time that passengers are actively boarding must equal the ratio of the passenger arrival rate λ_{ij} to the rate μ at which they board. The denominator $1 - \lambda_{ij}/(\overline{\eta}\lambda_{ij}^B)$ is the share of time a bus is present, which decreases with the arrival rate of busloads $\lambda_{ij}/\overline{\eta}$ of passengers relative to the bus arrival rate λ_{ij}^B. Second, inserting this probability into Equation (1) yields steady-state loading time,

$$L_{ij} = \frac{\overline{\eta}}{\lambda_{ij}} - \frac{1}{\lambda_{ij}^B}. \tag{2}$$

I next use three special cases to isolate key mechanisms. Case 1 ($\mu, \rho \to \infty$) isolates Market-Mohring increasing returns to demand. Case 2 ($\mu < \infty, \rho \to \infty$) adds boarding congestion while shutting down the gaps between successive buses. Case 3 ($\mu < \infty, \rho < \infty$) includes both frictions and introduces a role for Bus Supply Insurance.

$\mu, \rho \to \infty$ First, I consider infinite queueing and bus arrivals efficiencies. As $\mu \to \infty$, passengers board instantaneously, and $\rho \to \infty$ sends the bus arrival rate $\lambda_{ij}^B \to \infty$ so that the gaps in loading between successive buses vanish. Instant boarding means that queues cannot form, $Q_{ij} \to 0$, and loading time, from Equation (2), simply equals the expected time for a full busload of passengers to arrive, $L_{ij} = \overline{\eta}/\lambda_{ij}$. In this special case, total wait times,

$$Q_{ij} + L_{ij} = \frac{\overline{\eta}}{\lambda_{ij}},$$

fall log-linearly with demand λ_{ij}, exactly the *Market Mohring Effect* discussed in Section III. Higher demand raises the probability of a nonzero queue p_{ij}, so buses fill and depart faster, and wait times fall. An effective capacity constraint, as in the present case, therefore lends increasing returns to the transit technology — routes with more passengers must have more frequent departures — independently of market structure. This benchmark, however, abstracts from the passenger queues observed in many privatized shared transit systems.

$\mu < \infty, \rho \to \infty$ Second, I allow for congestion in boarding via a finite queueing efficiency $\mu < \infty$ but keep gaps between buses loading shut down ($\rho \to \infty$). This case collapses to a standard M/M/1 queue, as in Hillier and Lieberman (2024), such that

$$Q_{ij} = \frac{1}{\mu - \lambda_{ij}}. \tag{3}$$

A steady-state exists only when $\mu > \lambda_{ij}$, but even so, as demand λ_{ij} approaches μ, more frequent clusters of arrivals generate queues and raise expected queueing time.

$\mu < \infty, \rho < \infty$ Third, I introduce a role for bus supply by allowing for congestion in both passenger boarding ($\mu < \infty$) and bus arrivals ($\rho < \infty$). In Appendix A.2, I show that expected queueing time then satisfies

$$Q_{ij} = \frac{1}{\mu\left(1 - \frac{\lambda_{ij}}{\overline{\eta}\lambda^B_{ij}}\right) - \lambda_{ij}}.$$

Relative to Equation (3), the extra factor $\left[1 - \lambda_{ij}/(\overline{\eta}\lambda^B_{ij})\right]$ captures pauses in loading between successive buses. The slower bus arrivals λ^B_{ij}, the smaller the denominator, and the more strongly increases in λ_{ij} push up queueing time. Conversely, higher bus supply raises λ^B_{ij}, shortens inter-bus gaps, and dampens queue amplification — in other words, provides *Bus Supply Insurance* against demand shocks. Importantly, this new mechanism uniquely affects small-vehicle transit, $\overline{\eta} < \infty$, where multiple buses must arrive in quick succession to clear a surge in demand.[17]

Associations

On each origin-destination route ij, a single association acts as a constrained monopolist and makes two decisions in sequence. It first Nash-bargains the fare with the city government, then chooses bus supply conditional on that fare to maximize flow profits.

Profits

A minibus association's bus supply b_{ij} cycles continuously on route ij: they arrive at the origin station, load passengers for expected time L_{ij}, travel to their destination j in road-congestion-affected time T_{ij}, and return to the origin. The previously-introduced function $\mathrm{g}\left(b_{ij}, L_{ij}, T_{ij}; \rho\right)$ maps this operating cycle into the bus arrival rate λ^B_{ij} at the origin.[18] Crucially, an association can choose to serve a given quantity λ_{ij} demanded with a bus supply b_{ij} in excess of the minimum needed to carry expected demand. In this sense, b_{ij} can be thought of as "surplus" bus supply, more akin to service quality, that dampens queue build-up via the aforementioned Bus Supply Insurance Effect.

Associations, then, earn revenue from fares, denoted by τ_{ij}, and incur operations and labor costs. Each trip on a route costs χ_{ij} in operations costs, while drivers earn a fixed wage $\overline{\omega}$ per unit time.[19] Flow profits on route

[17] On a large-vehicle route, the resulting "queue" would disappear upon the arrival of the next metro or bus.

[18] My parameterization of $\mathrm{g}(\cdot;\rho)$ will account for travel time from j back to i; I implicitly assume that buses return empty, consistent with common practice in Cape Town.

[19] As discussed in Section II, fixed wages reflect the dominant driver payment modality in South Africa. The less-common

ij therefore equal the arrival rate of busloads times the per-trip net revenue in brackets, minus wage payments to the drivers of the b_{ij} buses in circulation:

$$\Pi_{ij} \equiv \underbrace{\frac{\lambda_{ij}}{\overline{\eta}}}_{\text{Busloads}} \left[\underbrace{\overline{\eta}\tau_{ij}}_{\text{Revenue}} - \underbrace{\chi_{ij}}_{\text{Ops. costs}} \right] - \underbrace{\overline{\omega} b_{ij}}_{\text{Driver wages}}. \tag{4}$$

I solve the association problem by backward induction, starting with bus supply.

Bus Supply

Conditional on bargained fares, associations choose bus supply b_{ij} to maximize profits, with first-order condition,

$$\left(\tau_{ij} - \frac{\chi_{ij}}{\overline{\eta}} \right) \frac{\partial \lambda_{ij}}{\partial b_{ij}} = \overline{\omega}. \tag{5}$$

The left-hand side of Equation (5) gives the marginal revenue from an additional bus, or the per-passenger margin times the demand response. To understand the latter, recall that additional bus supply increases the bus arrival rate λ_{ij}^{B}, reduces wait times through the Bus Supply Insurance Effect, and thereby increases demand λ_{ij}. The right-hand side reflects the marginal cost, the driver wage $\overline{\omega}$. Higher fares τ_{ij} increase the marginal revenue term, so associations internalize more of the value created by Bus Supply Insurance.

Fares

Before choosing bus supply, each route-level association Nash-bargains its fare with the city government, taking all other such pairwise route-level bargains as given. This "Nash-in-Nash" solution concept, standard in bilateral-oligopoly settings with spillovers (Collard-Wexler et al. (2019) and Yürükoğlu (2022)), provides a tractable approximation to the complex negotiations and enforcement actions which underlie Cape Town's government-mediated fare scheme. I define a Nash objective which trades off association profits Π_{ij}, as in Equation (4), against benevolent government motives, captured by the average commuter utility $\overline{\Omega}_{ij}$, conditional on home location i and work location j.[20] In particular, given association bargaining power β, fares τ_{ij} solve $\max_{\tau_{ij}} \left\{ \beta \log \Pi_{ij} + (1-\beta)\, \overline{\Omega}_{ij} \right\}$, with associated first-order condition

$$\frac{\beta}{\Pi_{ij}} \frac{\partial \Pi_{ij}}{\partial \tau_{ij}} + (1-\beta)\, \frac{\partial \overline{\Omega}_{ij}}{\partial \tau_{ij}} = 0.^{21} \tag{6}$$

commission-based pay could generate incentives less aligned with consumer welfare, encouraging drivers to drive erratically to maximize fare revenue in an echo of longstanding concerns surrounding civil servant incentives (Bandiera et al. (2024)). In my model, the utility costs later estimated for each mode could capture unsafe driving. Drivers paid on commission would certainly have additional incentive to depart only when full, as I assume in my model.

[20] Note that this average expected commuter utility by home and work location uses weights equal to a skill group's share of the aggregate city population.

[21] Note that I calculate this first-order condition using numerical derivatives of the Nash bargaining objective function.

The resulting fares rise with association bargaining power and converge to the monopoly benchmark, $\frac{\partial \Pi_{ij}}{\partial \tau_{ij}} = 0$, as $\beta \to 1$. Higher fares, recall, increase the marginal return to bus supply and thus strengthen associations' incentives to provide Bus Supply Insurance, but they also depress demand and therefore inhibit the realization of Market-Mohring increasing returns.

Commuters

The demand side of my model follows a standard quantitative spatial framework with exogenous amenities and wages. Commuters of skill g, either low or high, choose home location (i), work location (j), and mode (m), either minibus, public bus/train, or car, to maximize total utility, $\overline{U}^g_{ijm} \equiv \theta^g_i + U^g_{ijm} + \omega^g_j + \nu\varepsilon_{ijm}$. Utility comprises (i) an exogenous home amenity θ^g_i; (ii) a deterministic commute value U^g_{ijm}; (iii) a fixed workplace wage ω^g_j; and (iv) a Gumbel-distributed idiosyncratic preference ε_{ijm} with variance scaled by ν.[22]

The deterministic commute value U^g_{ijm} accounts for mode-specific utility costs, commute time, and monetary costs. I use a linear approximation to a micro-founded commute model, detailed in Online Appendix C.1. For minibus commuters,

$$U^g_{ijM} \equiv -\underset{\uparrow}{\kappa^g_M} - r\omega^g_j \Bigg(\underset{\uparrow}{Q_{ij}} + \underset{\uparrow}{L_{ij}} + \underset{\uparrow}{T_{ij}} \Bigg) - \underset{\uparrow}{\tau_{ij}}. \tag{7}$$

utility cost; queueing time; loading time; travel time; fare

The skill-specific utility cost κ^g_m captures non-pecuniary characteristics such as crime risk or driving safety. The next term accounts for time costs: commuters' disutility per minute of total commute time (in parentheses) equals the *product* of the time preference rate r and the wage ω^g_j. The final term subtracts off the minibus fare τ_{ij}. Utility on other modes follows the same structure. Public bus/train commuters incur a utility cost κ^g_P as well as exogenous fares τ_{ijP}, wait a fixed headway H_{ij}, and travel a fixed T_{ijP} minutes to their destination, so that $U^g_{ijP} \equiv -\kappa^g_P - r\omega^g_j \left(H_{ij} + T_{ijP}\right) - \tau_{ijP}$.[23] Car commuters incur a utility cost κ^g_A, fixed monetary cost τ_A, and have the same on-road, congestion-affected travel time T_{ij} as minibuses, so $U^g_{ijA} \equiv -\kappa^g_A - r\omega^g_j T_{ij} - \tau_A$.

Standard Gumbel aggregation implies that a skill-g commuter chooses to commute by mode m from home i to work j with probability

$$\pi^g_{ijm} = \frac{\exp\left(\theta^g_i + U^g_{ijm} + \omega^g_j\right)^{1/\nu}}{\sum_{i',j',m'} \exp\left[\theta^g_{i'} + U^g_{i'j'm'} + \omega^g_{j'}\right]^{1/\nu}}. \tag{8}$$

Lower ν implies stronger sensitivity to deterministic components such as commute time and fares. Minibus passenger arrivals then satisfy

$$\lambda_{ij} \equiv \sum_g N^g \pi^g_{ijM}, \tag{9}$$

[22] I do not allow workers to combine modes; in my stated preference survey, only 4.6% of minibus riders report also using public buses or trains over the course of their commutes, and 3.1% of minibus riders report also using a car or motorcycle.

[23] Note that, as detailed in Online Appendix D.4.5, I allow public commuters to walk to transit stops and transfer between buses and trains in my calibration of H_{ij} and T_{ijP}, such that a public route exists between any origin and destination.

such that local amenities and wages jointly shape which minibus routes grow busy enough to realize Market-Mohring reductions in wait times. Also by Gumbel properties, expected commuter utility across modes conditional on home, work, and skill satisfies $\overline{\Omega}_{ij}^{g} \equiv \nu \log\left[\sum_{m'} \exp\left(\theta_i^g + U_{ijm'}^g + \omega_j^g\right)^{1/\nu}\right]$. The route-level commuter utility in the fare bargaining problem associated with Equation (6) then averages $\overline{\Omega}_{ij}^{g}$ across skill groups using population-share weights: $\overline{\Omega}_{ij} \equiv \left(\sum_{g'} N^{g'}\right)^{-1} \sum_{g'} N^{g'} \overline{\Omega}_{ij}^{g'}$.

Road Congestion

I allow road congestion to affect car and minibus travel times through a simplified network. For tractability, I require that vehicles follow the path $s(i,j)$ that minimizes free-flow travel time. Vehicle inflow $x_{kk'}$ onto an individual network link kk' equals the sum of the commuter inflows, adjusted by vehicle capacity, over origin-destination pairs whose paths cross that link, i.e. $x_{kk'} = \sum_{i',j':kk' \in s(i',j')} \sum_{g'} N^{g'} \left(\pi_{i'j'M}^{g'}/\overline{\eta} + \pi_{i'j'A}^{g'}\right)$. Link-level travel time $t_{kk'}$ increases with vehicle inflow according to $t_{kk'} \equiv \bar{t}_{kk'} x_{kk'}^{\gamma}$, where $\bar{t}_{kk'}$ denotes free-flow travel time and γ is the road congestion elasticity. Finally, total travel time for minibuses and cars equals the sum of link-level travel times along $s(i,j)$, or $T_{ij} \equiv \sum_{kk' \in s(i,j)} t_{kk'}$. Because minibuses carry more passengers per vehicle, they impose a lower congestion externality per commuter than cars; shifts from cars to minibuses can therefore actually *reduce* road congestion.

Equilibrium and Efficiency

I now characterize equilibrium. A set of parameters and an exogenous geography jointly determine minibus supply, minibus fares, and commuter choice probabilities:

Definition. (Equilibrium) *Given parameters* $\{r, \nu, \boldsymbol{\kappa}, \tau_A, \boldsymbol{N}, \mu, \overline{\eta}, \rho, \overline{\omega}, \beta, \gamma\}$ *and the model geography* $\left\{\boldsymbol{\theta}, \boldsymbol{\omega}, \boldsymbol{\chi}, \bar{\boldsymbol{t}}, \boldsymbol{H}, \boldsymbol{T_P}, \boldsymbol{\tau_P}\right\}$*, an equilibrium is a vector* $\{\boldsymbol{b}, \boldsymbol{\tau}, \boldsymbol{\pi}\}$ *such that (i) associations maximize profits according to* (5)*; (ii) the Nash bargaining condition* (6) *holds; and (iii) commuter demand satisfies* (8)*.*

Efficiency

To characterize efficiency, I first define welfare and then the social planner's problem. The standard Gumbel properties again imply that ex-ante expected utility for skill group g satisfies $\overline{\Omega}^{g} \equiv \nu \log\left[\sum_{i',j',m'} \exp\left(\theta_{i'}^g + U_{i'j'm'}^g + \omega_{j'}^g\right)^{1/\nu}\right]$. I assume that commuters receive equal per-capita rebates of aggregate minibus profits, such that

Definition. (Welfare) Welfare*,* Ω*, equals the aggregate ex-ante expected utility of commuters plus aggregate minibus profits* $\Pi \equiv \sum_{i,j} \Pi_{ij}$*:*

$$\Omega \equiv \sum_{g} N^g \overline{\Omega}^g + \Pi. \tag{10}$$

The social planner then chooses allocations subject to the queueing technology, characterized by the numerically-simulated functions $Q\left(\lambda_{ij}, \lambda_{ij}^B\right)$ and $L\left(\lambda_{ij}, \lambda_{ij}^B\right)$, the bus arrivals function, and the road congestion technology.

Definition. *(Planning Problem) The social planner chooses bus supply* $\boldsymbol{b}$ *for each route as well as skill-group-specific commute choice probabilities* $\boldsymbol{\pi}$ *for each home, work, and mode to maximize welfare,* Ω*, as in* (10)*, subject to the queueing technology,* $Q_{ij} = Q\left(\lambda_{ij}, \lambda_{ij}^B\right)$ *and* $L_{ij} = L\left(\lambda_{ij}, \lambda_{ij}^B\right)$*, the bus arrivals function* $\lambda_{ij}^B = g\left(b_{ij}, L_{ij}, T_{ij}; \rho\right)$*, the road congestion technology* $T_{ij} = \sum_{kk' \in s(i,j)} \bar{t}_{kk'} x_{kk'}^{\gamma}$*, and the adding-up constraint* $\sum_{i,j,m} \pi_{ijm}^g = 1$ *for each skill group* g*.*

The model features three potential sources of inefficiency: (i) the Market Mohring Effect of minibus demand on wait times; (ii) the classic Road Congestion Externality; and (iii) the Bus Supply Insurance Effect on queue build-up. I now formally define the notion of internalization for each of these three effects. Intuitively, associations could, in theory, internalize the Market Mohring Effect by setting fares which induce optimal commuter demand. Road congestion naturally depends solely on the choices of car and minibus commuters, and associations, for their part, could directly internalize the Bus Supply Insurance Effect through their choice of b_{ij}. Note that, since the Market Mohring Effect and the Road Congestion Externality both operate through commuter demand, i.e. choice probabilities $\boldsymbol{\pi}$, I conceptually separate the two by holding either on-road travel times or minibus passenger arrival rates, respectively, fixed.

Definition. (Market Mohring Effect) *Given an equilibrium* $\{\boldsymbol{b}, \boldsymbol{\tau}, \boldsymbol{\pi}\}$*, associations internalize the Market Mohring Effect whenever demand* $\boldsymbol{\pi}$ *maximizes welfare,* $\Omega\left(\boldsymbol{b}, \cdot\right)$*, conditional on bus supply* b_{ij} *and on-road travel times* T_{ij} *and subject to* $Q_{ij} = Q\left(\lambda_{ij}, \lambda_{ij}^B\right)$*,* $L_{ij} = L\left(\lambda_{ij}, \lambda_{ij}^B\right)$*, the bus arrivals function* $\lambda_{ij}^B = g\left(b_{ij}, L_{ij}, T_{ij}; \rho\right)$*, and the adding-up constraint.*

Definition. (Road Congestion Externality) *Given an equilibrium* $\{\boldsymbol{b}, \boldsymbol{\tau}, \boldsymbol{\pi}\}$*, commuters internalize the Road Congestion Externality whenever minibus and car demand* $\{\boldsymbol{\pi_M}, \boldsymbol{\pi_A}\}$ *maximizes welfare,* $\Omega\left(\boldsymbol{b}, \cdot\right)$*, conditional on bus supply* b_{ij} *and minibus passenger arrival rates* λ_{ij} *and subject to* $Q_{ij} = Q\left(\lambda_{ij}, \lambda_{ij}^B\right)$*,* $L_{ij} = L\left(\lambda_{ij}, \lambda_{ij}^B\right)$*, the bus arrivals function* $\lambda_{ij}^B = g\left(b_{ij}, L_{ij}, T_{ij}; \rho\right)$*, the road congestion technology* $T_{ij} = \sum_{kk' \in s(i,j)} \bar{t}_{kk'} x_{kk'}^{\gamma}$*, and the adding-up constraint.*

Definition. (Bus Supply Insurance Effect) *Given an equilibrium* $\{\boldsymbol{b}, \boldsymbol{\tau}, \boldsymbol{\pi}\}$*, associations internalize the Bus Supply Insurance Effect whenever* $\boldsymbol{b}$ *maximizes welfare,* $\Omega\left(\cdot, \boldsymbol{\pi}\right)$*, conditional on commuter choice probabilities and subject to* $Q_{ij} = Q\left(\lambda_{ij}, \lambda_{ij}^B\right)$*,* $L_{ij} = L\left(\lambda_{ij}, \lambda_{ij}^B\right)$*, and the bus arrivals function* $\lambda_{ij}^B = g\left(b_{ij}, L_{ij}, T_{ij}; \rho\right)$*.*

However, Nash bargaining over fares gives associations a form of limited market power, so they induce optimal wait times only in a double-knife-edge case. In particular, two Hosios (1990)-type conditions require that bargained fares fully price in (i) the effect of demand on wait times *and* (ii) the marginal bus supply "insurance" value.[24] Formally,

Proposition 1. *Associations internalize the Market Mohring Effect if and only if*

$$\tau_{ij} = \frac{\chi_{ij}}{\overline{\eta}} + \sum_{g'} N^{g'} \pi_{ijM}^{g'} r \omega_j^{g'} \left[\frac{\partial Q_{ij}}{\partial \lambda_{ij}} + \frac{\partial L_{ij}}{\partial \lambda_{ij}} + \left(\frac{\partial Q_{ij}}{\partial \lambda_{ij}^B} + \frac{\partial L_{ij}}{\partial \lambda_{ij}^B} \right) \left. \frac{\partial \lambda_{ij}^B}{\partial \lambda_{ij}} \right|_{T_{ij}} \right] \tag{11}$$

[24] Naturally, absent appropriate road pricing, car and minibus commuters will not account for the road congestion spillovers they impose. Note that association bargaining power could in-theory induce minibus fares which exactly happen to price in road congestion, but, in my model, the same will never be true for car commuters.

and the Bus Supply Insurance Effect if and only if

$$\tau_{ij} = \frac{\chi_{ij}}{\overline{\eta}} - \left(\frac{\partial \lambda_{ij}}{\partial b_{ij}}\right)^{-1} \sum_{g'} N^{g'} \pi^{g'}_{ijM} r \omega^{g'}_{j} \left(\frac{\partial Q_{ij}}{\partial \lambda^{B}_{ij}} + \frac{\partial L_{ij}}{\partial \lambda^{B}_{ij}}\right) \frac{\partial \lambda^{B}_{ij}}{\partial b_{ij}}. \tag{12}$$

Commuters always fail to internalize the Road Congestion Externality.

Proof. See Appendix A.3. □

Generically, privatized shared transit market power is a double-edged sword. On the one hand, greater association bargaining power β raises fares, restricts the quantity λ_{ij} demanded, and—given the Market-Mohring increasing returns—raises wait times. On the other hand, for a given quantity demanded, higher bargaining power and thus fares raise the marginal revenue of additional bus supply *per passenger* in Equation (5), so associations provide better service quality, in the form of greater Bus Supply Insurance.[25] Thus, higher market power restricts demand-side increasing returns but strengthens supply-side insurance incentives.

Discussion

My model abstracts away from several margins discussed in the quantitative spatial literature. First, I do not model production and thus cannot endogenize wages or allow for agglomeration economies. My setting lacks the commercial floorspace and wage data necessary to calibrate a production function, as in Tsivanidis (2023). Moreover, Cape Town's recent transport investments very likely did not generate enough relocation to quantify agglomeration spillovers, and, in any case, the literature typically finds only moderate agglomeration elasticities (Combes and Gobillon (2015)). Nonetheless, in Online Appendix E.3, I extend the model to include agglomeration spillovers and re-calculate the social optimum using an agglomeration elasticity from the literature. Second, I abstract away from differences between home and work location choice elasticities, as in Tsivanidis (2023) or Zarate (2024). Here again, the South African context lacks exogenous variation to credibly estimate these parameters. Online Appendix E.3 nonetheless characterizes the social optimum under nested logit demand. Third, I focus on rush-hour work commutes. Resource constraints prevented the collection of station count data throughout the day, but my queueing and bus arrival frameworks could be integrated with state-of-the-art models of non-work trips (Miyauchi et al. (2022)). A suggestive low-demand simulation in Online Appendix E.3 indicates that off-peak times likely offer even greater scope for the realization of Market-Mohring increasing returns.

Similarly, my model of the minibus sector makes a number of simplifying assumptions. First, I do not allow associations or drivers to have their buses depart less-than-full. Though I could tractably allow associations to set the bus departure rate, I omit this margin because buses in my data almost always depart full. Online Appendix E.4 evaluates potential welfare gains from earlier, partially-full departures. A further extension could allow en-route pickup in origin neighborhoods, at the cost of longer travel time. Second,

[25] Note that, though higher β typically induces associations to choose closer-to-efficient bus supply b_{ij}, even in the case of $\beta \to 1$, associations will not set exactly efficient bus supply when there are multiple worker types due to a Spence (1975)-type quality distortion.

the model does not explicitly include all possible interactions across associations or between associations and government. Associations' supply choices correspond to a Cournot equilibrium, and route-level fare bargaining is Nash-in-Nash, conditional on other such bargains. However, governments could also contract on quality. In Online Appendix E.4, I allow the government to bargain over both fares and bus supply; because the government can then directly induce greater Bus Supply Insurance, the resulting allocation moves closer to the social optimum. Additional extensions could incorporate extra-legal interactions across associations.

Finally, the increasing returns in my model could potentially generate multiple equilibria. Road and bus-arrival congestion could nonetheless offset these increasing returns and restore uniqueness. While I cannot rule out multiplicity analytically, I initialize my solution algorithm at lower and higher values of λ_{ij} and b_{ij}, relative to equilibrium; in each case, the algorithm converges to virtually the same equilibrium.[26]

V. Estimation of Minibus and Demand Parameters

I estimate the model's structural parameters in four main steps. First, I devise instrumental variables strategies which employ the station counts of passengers and buses to identify queueing efficiency μ and arrival efficiency ρ. Second, I estimate mode-specific utility costs κ_m^g, the rate of time preference r, and the Gumbel scale ν from stated preferences. Third, I estimate the road congestion elasticity from TomTom traffic data and externally calibrate my model to $I = 18$ *transport analysis zones* in Cape Town. Fourth, conditional on other parameters, I internally calibrate the association bargaining power β and driver wage $\overline{\omega}$ and then invert the model to obtain home location-specific amenities θ_i^g and work location-specific wages ω_j^g. Table 1 summarizes all calibrated parameters.

Minibus Loading

First, I estimate two key parameters of the minibus loading process: passenger queueing efficiency μ and bus arrival efficiency ρ. I derive each estimating equation from an empirical version of my queueing framework in Online Appendix C.2 which incorporates (i) unobserved loading delays, which begin whenever a bus arrives and end at Poisson rate ι_{ij}, and (ii) an unobserved bus arrival speed, denoted by ς_{ij}.

Queueing Efficiency μ I estimate queueing efficiency using minibus loading times. Were there always a queue of passengers and no interruptions to loading, I could estimate queueing efficiency directly from the average loading time, which, since each passenger takes an average time $1/\mu$ to board, would satisfy $L_{ij} = \overline{\eta}/\mu$. However, this loading process may periodically stop when the queue empties of passengers or due to unobserved weather or other shocks. In either case, the total loading time is no longer directly proportional to μ.

Instead, Equation (1), translated to the route ij by five-minute time period t by date d level, implies that

[26] In particular, I initialize the algorithm at two alternative starting points: (i) at 50% lower λ_{ij} and b_{ij} on each route, as compared to the baseline equilibrium, and (ii) at 50% higher λ_{ij} and 25% higher b_{ij}. In no case does the median absolute deviation over routes between the solution reached from the new starting point and the baseline calculated equilibrium bus supply b_{ij}, passenger arrivals λ_{ij}, or fares τ_{ij}, taken as a proportion of the baseline equilibrium value, exceed 6×10^{-4}.

Table 1. Calibrated Parameters

Parameter	Description	Value
Externally Calibrated		
I	Number Locations	18
N^g	Commuter Populations	
$\bar{t}_{ij}$	Free-Flow Driving Time	
H_{ij}	Public Bus/Train Wait Time	
T_{ijP}	Public Bus/Train Travel Time	
τ_A	Car Commute Cost	12.4
τ_{ijP}	Public Bus/Train Fare	
$\overline{\eta}$	Minibus Capacity	15
χ_{ij}	Minibus Operating Cost	
Minibus Loading		
μ	Queueing Efficiency	3.54
ρ	Arrival Efficiency	0.18
Road Congestion		
γ	Road Congestion Elasticity	0.09

Parameter	Description	Value
Stated Preference		
r	Commuter Rate of Time Pref.	0.001
ν	Gumbel Shape	4.76
κ_M^l	Low-Skill Minibus Util. Cost	7.7
κ_M^h	High-Skill Minibus Util. Cost	15
κ_P^l	Low-Skill Public Util. Cost	3.6
κ_P^h	High-Skill Public Util. Cost	9.2
Internally Calibrated		
β	Assoc. Bargaining Power	0.19
$\overline{\omega}$	Driver Wage	0.0007
Model Inversion		
θ_i^g	Amenities	
ω_j^g	Wages	

Notes: This table displays the full set of estimated model parameters. The externally calibrated parameters and geography come primarily from the 2013 Cape Town Household Travel Survey as well as the Azure API. The minibus loading parameters are estimated using the station counts. The road congestion elasticity uses TomTom API data. The stated preference estimation uses my new survey and an existing module from the aforementioned 2013 survey. The internal calibration minimizes the distance to moments in minibus on-board tracking data and the station counts, and the model inversion again employs the 2013 survey data.

expected bus loading times depend on the probability p_{ijtd} that a nonzero queue of passengers forms, conditional on a bus being present, or "queue prevalence" for short. To calculate p_{ijtd} in the data, I restrict attention to the every-five-minute queue length observations within a rolling window around t during which a bus is present in the station and then calculate the share of these observations with nonzero queue length.[27] Intuitively, the more often a queue of passengers forms, the less time buses will spend waiting for passengers and the lower L_{ij}. The faster any queueing passengers board, i.e. the higher μ, the more strongly the mere presence of a queue will shorten loading times. With queue prevalence in hand, I then regress the loading time L_{sijtd} of bus s on route ij departing during five-minute period t on date d on queue prevalence p_{ijtd} according to

$$L_{sijtd} = \frac{\overline{\eta}}{\mu}\left(p_{ijtd}\right)^{-1} + \underbrace{\delta_{ijd} + \delta_{itd} + \zeta_{sijtd}}_{\equiv \varepsilon_{sijtd}}. \tag{13}$$

I parameterize the unobservable term ε_{sijtd}, which contains the unobserved rate ι_{sijtd} at which loading

[27] Specifically, for each route ij as well as five-minute time period t and date d corresponding to a bus's departure, I calculate, within a one-hour rolling window centered on t, the share of five-minute periods at the start of which I see one or more passengers in the queue *and* a bus present in the loading area. To obtain the conditional probability p_{ijtd}, I then divide by the total share of time periods t within this hour window at the start of which a bus is present. By way of example, an hour window with a bus in the loading area for route ij during three of the six five-minute time periods and no passengers queueing during two of those three would yield $p_{ijtd} = \frac{1}{3}$.

delays end, as the sum of route-by-date and origin-by-five minute time period-by-date fixed effects as well as an idiosyncratic shock.[28]

The aforementioned unobserved interruptions of the usual loading process, e.g. weather or special events, pose the primary threat to identification. The direction of bias is ex-ante ambiguous: disruptions which coincide with higher demand and thus more frequent queues would bias μ towards zero, while shocks such as rain, which might directly depress passenger arrivals, could bias μ upwards. Note that route-by-date fixed effects, δ_{ijd}, capture time-invariant causes of interruptions on specific routes. Furthermore, origin-by-time-by-date fixed effects, δ_{itd}, absorb any time-variant delays common to all routes from a given origin, such as station infrastructure-related constraints which bind only at the busiest times. Thus, the error term ζ_{sijtd} contains only time-varying, route-specific factors within a given origin station that might delay or stop loading.

I address any remaining endogeneity with an instrument for queue prevalence: the number of commuters from i to j who leave for work during a rolling one-hour window around t, measured in a 2013 survey.[29] Intuitively, higher commuter numbers should translate directly into the kind of spikes in demand which increase the prevalence of queues. This distribution of work start times instrument satisfies the exclusion restriction, provided it does not vary systematically with idiosyncratic interruptions to loading. Since I measure queue prevalence and bus loading times in 2022, such minute-by-minute trends in interruptions would have to persist over nine years to violate instrumental exogeneity. Even so, commute start times remain exogenous, provided the timing of weather or events which affect specific routes differentially within a given origin does not systematically vary with the corresponding commuters' start times. In Online Appendix D.1.2, I indeed find little evidence of hour-by-hour trends in the variances of weather and traffic speeds.

Table 2 displays the queueing efficiency values implied by instrumental variables estimation of Equation (13).[30] I start with my baseline specification in Column 1; in Columns 2 and 3, I add route-by-date and origin-by-five minute time period-by-date fixed effects, respectively. The efficiency $\hat{\mu} = 3.54$ in my preferred specification in Column 3 indicates that the average commuter takes about 17 seconds to board a minibus.[31] The literature offers little precedent for comparison, but this short boarding time suggests that most minibus routes could easily absorb additional passenger arrivals without an attendant build-up of queues.

In Online Appendix D.1.3, I assess the robustness of my estimated queueing efficiency using alternative instruments and an alternative estimator. The instruments include (i) an interaction of the baseline work-start-times instrument with an indicator for MyCiti BRT-affected origins and (ii) within-hour variation in commuter work start times, which allows me to control for demand levels and strengthens the exclusion restriction. I also estimate μ from average loading times for buses that arrive to at least a full busload of waiting passengers. Though the confidence intervals of these alternative estimates always come close to including the baseline

[28] Given the number of routes in my sample, I lack the variation to additionally include destination-by-five minute time period-by-date fixed effects, δ_{jtd}. Furthermore, note that all routes starting at a given origin i were surveyed only on one date, so that the d subscript is redundant, but I nevertheless index by both time and date for clarity.

[29] At the time of writing, the 2013 Cape Town Household Travel Survey offered the most recent available spatially-granular data on commuter demand.

[30] I relegate the first stages to Online Appendix Table D.1.

[31] Note that I obtain the implied value of μ by taking the reciprocal of the coefficient on (inverse) queue prevalence in Equation (13) and multiplying by the externally-calibrated minibus capacity $\overline{\eta}$.

Table 2. Queueing Efficiency Estimates

Parameter	(1) Bus loading time	(2) Bus loading time	(3) Bus loading time
$\overline{\eta}/\mu$	3.49	1.94	4.24
Minibus Capacity/Queueing Efficiency	(0.93)	(0.45)	(1.78)
Implied μ	4.30	7.75	3.54
Queueing Efficiency	(1.15)	(1.80)	(1.48)
Route-by-Date FE		✓	✓
Origin-by-Time-by-Date FE			✓
Observations	1077	1077	1023
First-Stage F Statistic	26.05	69.37	12.39

Notes: Robust standard errors in parentheses, clustered at the route-by-five minute time period-by-date level. Table presents estimates of (13) over bus departures in my station count data, with fixed effects included, as noted, for route-by-date and origin-by-five minute time period-by-date. I instrument for queue prevalence, i.e. the probability of a nonzero passenger queue conditional on a bus being present in the loading area in the rolling hour centered around a bus's departure, using the log number of commuters living in the transport analysis zone (TAZ) where the route originates and working in the route's destination TAZ who report leaving their home during this same rolling hour, calculated from the 2013 Cape Town Household Travel Survey. I additionally report the effective first-stage F statistic following Olea and Pflueger (2013). Finally, I obtain the implied value of μ by taking the reciprocal of the coefficient on (inverse) queue prevalence and multiplying by the externally-calibrated minibus capacity $\overline{\eta}$.

estimate, I also later subject my main results to a sensitivity analysis.

Bus Arrival Efficiency ρ For empirical implementation, I impose a functional form for the bus arrival rate function consistent with a stylized model of bus operations, such that $\mathrm{g}\left(b_{ij}, L_{ij}, T_{ij}; \rho\right) \equiv \rho/\log\left\{\exp\left[\rho\left(\frac{2T_{ij}}{b_{ij}} - L_{ij}\right)\right] + 1\right\}$.[32] In the station count data, I then estimate an equation for the expectation of the gap_{sijhd} in minutes between the departure of the preceding bus from the origin station and the arrival of bus s on route ij during hour h on date d. Given Poisson arrivals at rate $\lambda_{ij}^{B} \equiv \mathrm{g}\left(b_{ij}, L_{ij}, T_{ij}; \rho\right)$, this conditional expectation decreases with the number of buses b_{ijhd}, more so, the greater arrival efficiency:

$$\mathrm{gap}_{sijhd} = \frac{1}{\rho}\log\left\{\exp\left[\rho\left(\frac{2T_{ij}}{b_{ijhd}} - L_{sijhd}\right)\right] + 1\right\} + \underbrace{\Gamma_{n(i)hd} + \Gamma_{n(j)} + \vartheta_{sijhd}}_{\equiv -\frac{1}{\rho}\log \varsigma_{sijhd}}. \tag{14}$$

The error term corresponds to a transformation of the unobserved arrival speed ς_{sijhd}, introduced in Online Appendix C.2. Intuitively, ς_{sijhd} captures factors such as neighborhood foot and car traffic and is therefore likely correlated with associations' bus supply choices. I absorb such congestion with origin transport

[32] Suppose that buses depart at equally-spaced intervals, return to the origin empty after dropping off their passengers at the destination, and experience no additional congestion in and around the minibus station. Given travel time T_{ij} on a route, a round trip would then take $2T_{ij}$, and one full "cycle" on the route would take $L_{ij} + 2T_{ij}$ minutes. To ensure equal intervals between bus arrivals, the next bus must arrive at the origin $\frac{L_{ij}+2T_{ij}}{b_{ij}}$ minutes after the first bus arrives and begins loading. At that moment, $\frac{L_{ij}+2T_{ij}}{b_{ij}} - L_{ij}$ minutes will have elapsed since the first bus finished loading and departed. For simplicity and motivated by the relative magnitudes in my data, I assume that L_{ij} is small relative to b_{ij}, such that $\frac{L_{ij}}{b_{ij}} \approx 0$ and the bus gap in a world without congestion in and around minibus stations approximately equals $\frac{2T_{ij}}{b_{ij}} - L_{ij}$. The full functional form of g then incorporates congestion in and around minibus stations, controlled by the arrival efficiency ρ.

analysis zone (TAZ) $n(i)$-by-hour-by-date and destination TAZ $n(j)$ fixed effects. The remaining threat to identification comes from route-specific idiosyncratic slowdowns in ϑ_{sijhd}.

I therefore estimate Equation (14) via GMM and instrument for the first term on the right-hand side with log origin-destination distance. Route length determines operations costs and thus bus supply but should not vary systematically with the degree to which, say, foot traffic affects a particular route within a given origin neighborhood. In my preferred GMM specification, I estimate an arrival efficiency of $\hat{\rho} = 0.178$.[33] The implied congestion in bus arrivals means that the gap between buses on the average route will never fall below around two minutes.

Stated Preferences

Next, I estimate key demand parameters, including commuters' value of time and preferences over modes. While the literature typically employs revealed preference approaches based on travel diary or commuter surveys (Tsivanidis (2023) and Zarate (2024)), I instead use stated preferences for two reasons. First, they allow me to value attributes that are difficult to systematically observe in practice, such as minibus station security guard presence, and will thus later help decompose commuters' preferences for minibuses. Second, stated preferences provide exogenous variation in commute attributes such as travel time. In contrast, in revealed preference data, observable attributes of different modes might vary systematically with unobserved comfort or crowding, and few natural instruments exist.

Using this exogenous variation, I can directly estimate the model's logit demand system via standard maximum likelihood. The attributes in the combined stated preference surveys correspond to the model in Section IV, with one notable addition: each alternative l in my own survey featured a set $\mathscr{Z}(l)$ of "quality improvements." I thus incorporate a skill-specific linear effect ξ_z^g of each improvement z on mode utility costs into Equation (8) and denote total wait time by H_l. Conditional on home and work locations, a skill-group-g respondent i chooses alternative l of mode $m(l)$ in a given choice set with model-implied probability

$$\pi_{il}^g = \frac{\exp\left[-\kappa_{m(l)}^g - \sum_{z\in\mathscr{Z}(l)} \xi_z^g - r\omega_i\left(H_l + T_l\right) - \tau_l + rH_l\tau_l\right]^{1/\nu}}{\sum_{l'} \exp\left[-\kappa_{m(l')}^g - \sum_{z\in\mathscr{Z}(l')} \xi_z^g - r\omega_i\left(H_{l'} + T_{l'}\right) - \tau_{l'} + rH_{l'}\tau_{l'}\right]^{1/\nu}}.^{34} \tag{15}$$

Differences in mode shares, all other attributes equal, identify relative utility costs, κ_m^g.[35] The sensitivity of respondents' choices to the presence of quality improvements identifies their effects, ξ_z^g, on utility costs. Crucially, the extent to which wait as well as travel time, T_l, differentially decrease the choice probabilities of respondents with higher income ω_i helps quantify the rate of time preference, r. Finally, the variation in fares, τ_l, translates into the Gumbel scale ν and allows me to calculate dollar willingness to pay for each attribute.

[33] Robust standard error = 0.09, $N = 1,158$ bus departures. I present full results in Online Appendix D.2 and, consistent with my model, restrict the mean of each set of fixed effects to equal zero.

[34] Unlike in the main text, I include a final higher-order term–the product of fares and wait time–as implied by the micro-founded model in Online Appendix C.1.

[35] While my own survey over-sampled minibus commuters, my survey included only minibus alternatives and so does not contribute to the identification of the relative utility costs across modes.

The left panel of Table 3 shows the resulting estimates, among the first from a developing-country context. My estimated rate of time preference, $\hat{r} = 0.001$, implies a value of time equal to 24% of the hourly wage, at the lower end of the 20-140% range of existing high-income-country estimates (Small and Verhoef (2007), Almagro et al. (2024), Buchholz et al. (2024), and Goldszmidt et al. (2020)). This value, low even relative to existing estimates for India (Kreindler (2022)), might reflect more limited housework and childcare commitments or a lack of available leisure opportunities in township neighbourhoods. Next, I estimate a Gumbel scale parameter $\hat{\nu} = 4.76$. The implied minibus own-price elasticity of demand, at -0.17, falls substantially below the elasticities of -0.5 to -0.7 estimated for public transit and ride-hailing in the US (Goldszmidt et al. (2020) and Almagro et al. (2024)) and suggests a limited availability of close substitutes.[36]

The remaining parameters in Table 3 characterize preferences over modes. The low- and high-skill minibus utility costs, $\hat{\kappa}^l_M = 7.68$ and $\hat{\kappa}^h_M = 15.03$, imply that both groups dislike minibuses, relative to public buses and trains as well as cars. In contrast, commuters in Latin America tend to prefer minibuses to publicly-provided bus rapid transit (Tsivanidis (2023) and Zarate (2024)). The right panel shows that all three quality improvements in my survey—station security guards, overloading bans, and speed-limit enforcement—significantly reduce minibus utility costs for low-skill commuters. The high-skill, for their part, would pay a full \$2.75 per commute for station security. The annual equivalent of this value exceeds women's already sizable willingness to pay for safer walking routes in India in Borker (2021) by up to a factor of five—unsurprising, perhaps, given Cape Town's violent crime rates.

To assess external validity, I examine selection into my stated-preference sample in Online Appendix D.3. Since the stated preference samples match the population on every observable except actual reported commute mode, I confirm that the estimates change little when I weight my own survey by realized commute mode choices. I then estimate mode-specific own-price elasticities separately using (i) the combined baseline stated preference sample, including the representative city-run survey, and (ii) my survey alone. The similarity between both sets of elasticities provides some reassurance that my minibus-commuter-heavy sample does not differ on unobservables, e.g. individual values of time, in ways that materially affect demand estimates. Consistent with this interpretation, my sample contains representative shares of groups typically estimated to have higher values of time, such as college-educated workers and women (Borghorst et al. (2021)).[37] As an additional test for selection, I compare estimates by interview location in Online Appendix D.3; the estimates remain broadly similar across respondents interviewed at minibus stations and elsewhere. I nonetheless find noisy evidence of a lower security value at one minibus station; to the extent that in-station respondents value security less, my estimates might indicate a lower bound.

I finally return to the two central challenges in stated preference estimation discussed in Section II: comprehension and hypothetical bias. Imperfect comprehension would add noise and likely attenuate estimates; correspondingly, the large and relatively precise effects I estimate for the three quality-improvement

[36] In contrast, Cape Town drivers, with an implied own-price elasticity of -1.62, substitute readily towards other modes. Note that, to calculate demand elasticities, I use median minibus fares from my on-board tracking data as well as the median calibrated public bus/train fare and calibrated car per-commute cost. Mode shares come from the 2013 Cape Town Household Travel Survey.

[37] See Online Appendix Table B.3. Note also that no respondent in my survey broke off the questionnaire before finishing all discrete choice experiments, which suggests that the time burden did not significantly impede respondents' commutes.

Table 3. Stated Preference Survey Estimates

Parameter	Value	
r	.001	
Commuter Rate of Time Pref.	(.0004)	
ν	4.76	
Gumbel Scale	(1.26)	
Utility Costs	*Low-Skill*	*High-Skill*
κ_P	3.63	9.17
Public Bus/Train Utility Cost	(0.51)	(1.89)
κ_M	7.68	15.03
Minibus Utility Cost	(1.56)	(3.55)

Parameter	Value	
Effects on Utility Costs	*Low-Skill*	*High-Skill*
ξ_{security}	-1.09	-2.75
Station Security	(0.390)	(0.84)
$\xi_{\text{no overloading}}$	-1.38	-1.39
Overloading Ban	(0.437)	(0.543)
$\xi_{\text{follows speed limit}}$	-1.36	-0.825
Speed Limit Enforcement	(0.445)	(0.465)

Notes: Robust standard errors in parentheses. Estimates reflect $N = 19,712$ individuals by choice sets by alternatives in either my newly-collected minibus stated preference survey (4,130 individuals by choice sets by alternatives, 413 unique individuals) or a stated preference module of the 2013 Cape Town Household Travel Survey (15,582 individuals by choice sets by alternatives, 407 unique individuals). The estimated parameters come from a multinomial logit model with choice probabilities given by (15). I normalize $\kappa_A^g = 0$ and restrict the sample to individuals employed outside the home between 25 and 65 years of age.

attributes provide at least suggestive evidence that respondents understood the task. In discrete choice experiments with multiple dimensions of costs, hypothetical bias in *relative* valuations cannot be conclusively signed. To nonetheless verify the plausibility of my estimates, in Online Appendix D.3.3, I re-estimate demand using revealed-preference commute choices from the 2013 Cape Town Household Travel Survey. Although this approach does not precisely identify the Gumbel scale ν, it yields a similar rate of time preference and high-skill dislike for minibuses and public modes. As a final over-identification check, Section VI characterizes respondents' ability to predict their actual commute modes.

External Calibration

I externally calibrate secondary parameters along with the model geography. I estimate the road congestion elasticity γ with data from TomTom's Traffic Stats API on traffic volume x_{ih} and travel time t_{ih} across road segments i in Cape Town and hours h. My model of road congestion implies a log-linear congestion function, $\log t_{ih} = \bar{t}_i + \gamma \log x_{ih} + \epsilon_{ih}$, where the unobservable ϵ_{ih} would include disruptions such as weather and special events (Barwick et al. 2022). In Column 1 of Online Appendix Table D.9, I obtain $\hat{\gamma} = \texttt{0.0917}$. Next, I directly obtain the car commute cost τ_A, minibus capacity $\overline{\eta}$, and the model geography from the data. The latter includes commute populations N^g, minibus operating costs χ_{ij}, free-flow driving times $\bar{t}_{kk'}$, and public bus/train wait and travel times $\{H_{ij}, T_{ijP}\}$ as well as fares τ_{ijP}.[38] I provide additional details in Online Appendix D.4.

[38] Since fixed costs typically dominate the variable costs of car ownership, I calibrate a single τ_A, constant across origins and destinations, based on estimates of the total car ownership cost per (half) day, as discussed in Online Appendix D.4. I calculate public bus/train fares τ_{ijP} directly from Cape Town's distance-based public transit fare scheme.

Table 4. Internal Calibration

Moment			Parameter		
Description	*Data*	*Model*	*Description*		*Value*
Median Minibus Fare ($)	1.05	1.05	β	Association Bargaining Power	0.19
Median Queue Length (Number Passengers)	4	3.99	$\overline{\omega}$	Minibus Driver Wage	0.0007

Notes: This table displays the moments used in internal calibration: the median minibus fare across routes in my onboard tracking data and the median number of passengers waiting in the queue across routes and five-minute periods in the station count data. In the model, I calculate medians across routes. I also list the model parameter heuristically corresponding to each moment, along with its internally-calibrated value.

Internal Calibration and Inversion

Conditional on all other parameters, I then match aggregate moments in my minibus data to obtain the remaining supply parameters and invert the model to obtain location-specific amenities and wages. Specifically, I match (i) the median minibus fare, τ_{ij}, across routes in the onboard tracking data, and (ii) the median queue length n_{ij}, in passengers, across routes and five-minute periods in the station counts. As in standard spatial models, I then exactly match each location's observed employment by residence and workplace to calibrate home location amenities θ_i^g and workplace wages ω_j^g, respectively.[39] The identification of these parameters thus relies on the accuracy of all other parameters. The second and third columns of Table 4 list the values of each moment in the data and across routes in my calibrated model. Heuristically, fares identify association bargaining power β, and queue length identifies the minibus driver wage $\overline{\omega}$. The final column lists the calibrated parameter values. Importantly, the modest association bargaining power reveals that the city government effectively caps fares far short of monopoly levels, contrary to popular perceptions of association market power.

VI. Model-Predicted Minibus Supply and Demand

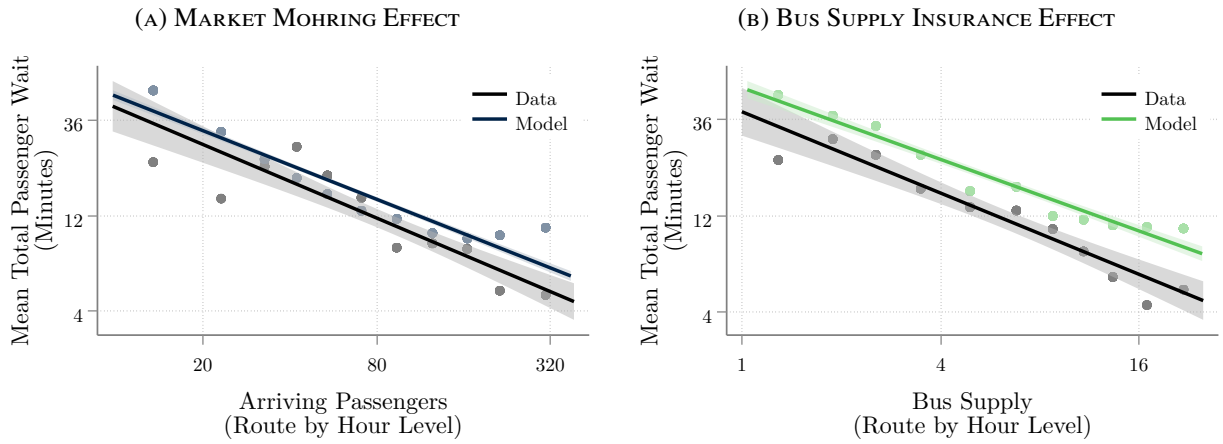
I now demonstrate that my model matches key moments: mode shares and fares, the Market Mohring and Bus Supply Insurance Effects, the spatial distribution of minibus supply and wait times, and the actual commute modes of stated preference respondents.

First, I assess how well the model predicts empirical mode shares, minibus fares, and the minibus network. In Online Appendix Tables E.1-E.2, I regress empirical outcomes on model predictions for minibus and car mode shares and minibus fares by origin and destination. Coefficients near one indicate that the model captures the core mechanisms governing mode shares and fares, idiosyncratic factors outside the model notwithstanding. Likewise, Online Appendix Figures A.4a-A.4b show that the model-predicted bus supply across origin-destination pairs mirrors the observed number of distinct minibus routes.[40] Importantly, the

[39] I measure employment by residence and workplace in the 2013 Cape Town Household Travel Survey and normalize, for each skill group, (i) the amenity of one location to zero and (ii) the average wage to the empirical skill-group average.

[40] Note that, in reality, unlike in my model, many neighborhood (transport analysis zone) pairs are linked by multiple distinct minibus routes.

FIGURE 8. WAIT TIMES IN DATA VERSUS MODEL



Notes: Panel (A) displays binned scatterplots and best-fit lines of the log-scale relationship between expected total passenger wait time, $Q_{ij} + L_{ij}$, and newly-arriving passengers per hour, proportional to λ_{ij}, across routes and hours in the station count data and across routes as predicted by the model. Panel (B) instead displays the relationship between expected total passenger wait time and bus supply b_{ij}.

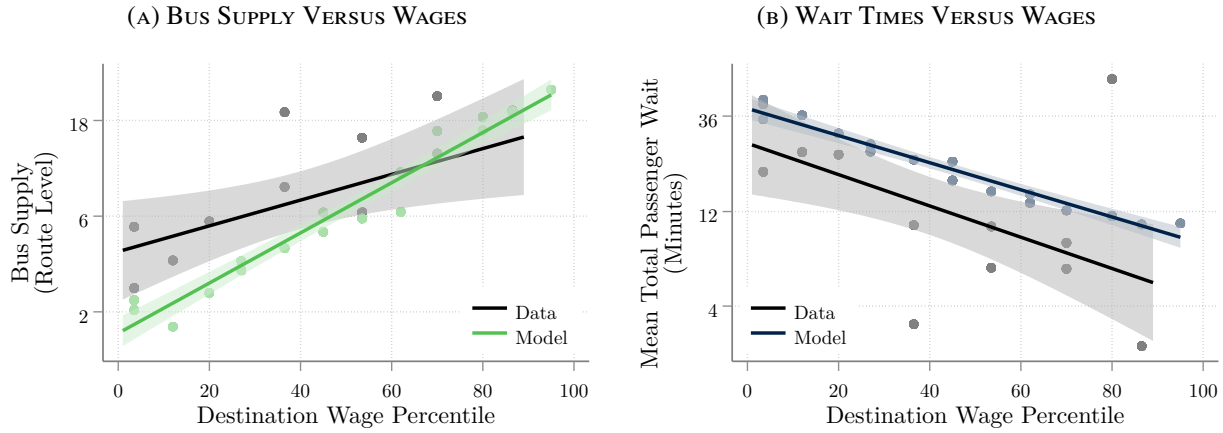
model matches the concentration of minibuses in central neighborhoods and the direct links between outlying suburbs. Consistent with this fit, Online Appendix Figure A.4c shows that the model matches the empirical relationship between passenger arrivals and bus supply.

Second, the model replicates both the targeted Market Mohring Effect and the non-targeted Bus Supply Insurance Effect. Figures 8a and 8b show binned scatterplots of total wait time versus passenger arrivals and bus supply, respectively, in the data in black and across model routes in color. Because Equation (13) effectively targets the Market Mohring Effect, the model successfully replicates the decline in wait times with passenger demand in Figure 8a. More importantly, it also matches the non-targeted Bus Supply Insurance relationship between bus supply and waits in Figure 8b.

Third, the model also matches key cross-route patterns in bus supply and wait times. I use average destination wages as a proxy for a route's attractiveness to commuters and, in Figures 9a-9b, plot binned route-level bus supply and wait times, respectively, against destination wage percentiles in data and model. My association supply model evidently captures the non-targeted pattern of greater Bus Supply Insurance on routes to higher-wage destinations, which attract more passengers. Even though queueing estimation does not use wages, wait times decline with destination wages by a similar magnitude in data and model, a further validation of the model-generated Market Mohring Effect.

Fourth, to assess the credibility of my stated-preference data, I compare respondents' reported "real-life" commute modes—unused in estimation—to those predicted by my model. This comparison provides an indirect test for the hypothetical bias discussed in Section II. The bars in Figure 10 indicate the shares of respondents who commute via each mode in reality (light colors) and as predicted by the model (dark colors). Respondents' actual *revealed* commute choices mirror model predictions, based on their hypothetical *stated* choices, quite closely. In particular, the model reproduces both the skill gradient in minibus use and the

FIGURE 9. BUS SUPPLY AND WAIT TIMES ACROSS ROUTES IN DATA VERSUS MODEL



Notes: Panel (A) displays binned scatterplots and best-fit lines of the relationship between route-level bus supply, b_{ij}, on a log scale, and the percentile of the average wage at the route's destination, separately by origin-destination transport analysis zone (TAZ) in the data and over routes as predicted by the model. Panel (B) instead displays, on the vertical axis, the route-level expected total passenger wait time, $Q_{ij} + L_{ij}$, on a log scale, again versus destination wage percentiles. In the data, I calculate average wait times and average total hourly bus supply by origin-destination TAZ pair using the station counts, and mean personal income by destination TAZ using the 2013 Cape Town Household Travel Survey. In the model, origin-destination TAZ pairs are synonymous with routes.

predominance of driving among high-skill commuters, largely through utility costs. Thus, even though respondents did not face the actual monetary or time costs associated with their stated choices, those choices appear informative about real-world preferences. As a final plausibility check, Online Appendix E.2 shows intuitive demographic heterogeneity in values of time and quality improvements—for example, women place higher value on time savings, consistent with Borghorst et al. (2021).

VII. EFFICIENCY OF PRIVATIZED TRANSIT PROVISION

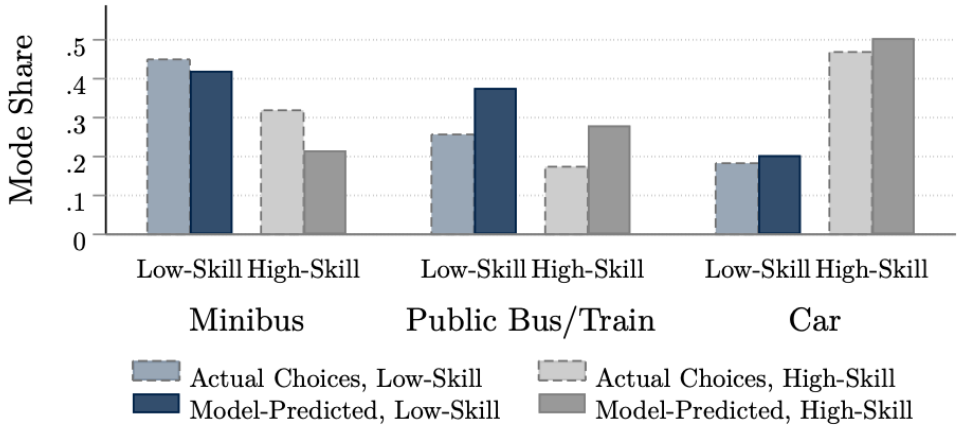
Finally, I return to my central question: how inefficiently does Cape Town's privatized shared transit network provide the "public" good of urban transit? I characterize the socially-optimal minibus demand and supply and then conclude with a decomposition of the determinants of minibus demand.

Social Planner Optimum

I now solve the planning problem defined in Section IV and then characterize the socially-optimal commuter choices $\boldsymbol{\pi}$ and minibus entry $\boldsymbol{b}$. In Table 5, I summarize the associated changes in aggregate mode shares and in the welfare measure Ω defined in Equation (10). I measure welfare changes as equivalent variation: the proportionate change in a skill group's wages ω_j^g, at baseline values of $\{\boldsymbol{b}, \boldsymbol{\tau}, \boldsymbol{Q}, \boldsymbol{L}, \boldsymbol{T}, \boldsymbol{\kappa}, \overline{\eta}, \boldsymbol{H}, \boldsymbol{T_P}\}$, that leaves the average commuter equally well off as in the counterfactual.[41]

[41]To calculate baseline equilibrium welfare at the skill-group level in a manner unaffected by minibus fares, I rebate minibus profits as follows: for each route, I multiply route-level profits by a group's share among minibus commuters on that origin-destination and then sum across routes (i.e. origin-destination pairs). At the social optimum, I allocate minibus operations and driver wage costs on a given route to each skill group according to their share of minibus commuters on that route, sum across all routes, and subtract the total from that skill group's total ex-ante expected commuter utility $N^g \overline{\Omega}^g$.

FIGURE 10. STATED PREFERENCE RESPONDENTS' ACTUAL VERSUS MODEL-PREDICTED COMMUTE MODES



Notes: This figure displays the shares of low- (non-college) and high-skill stated preference respondents in my newly-collected minibus stated preference survey or a stated preference module of the 2013 Cape Town Household Travel Survey who report that they commute via each mode, alongside the model-predicted shares. The latter predictions make use of regressions of two outcomes, cost and commute time, on age, gender, education, and income. I run regressions separately for each outcome and mode in the 2013 (revealed preference) household data. These regressions allow prediction of respondents' hypothetical commute time and cost via each mode based on the aforementioned demographic characteristics. The estimated demand parameters, together with these predicted times and costs, imply choice probabilities π^g_{ijm}; the model predictions are sums of these choice probabilities within each skill and mode.

TABLE 5. SOCIAL PLANNER OPTIMUM VERSUS STATUS QUO EQUILIBRIUM

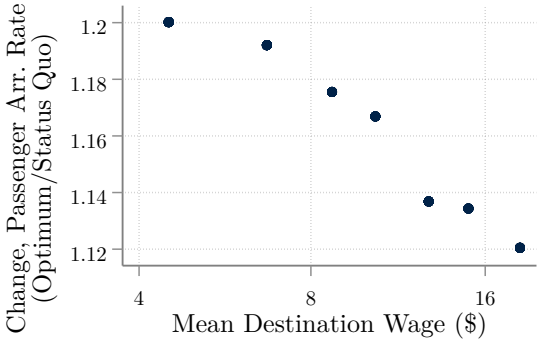
	Optimal b_{ij} Only		Optimal π^g_{ijm} Only		Optimal b_{ij} and π^g_{ijm}	
Δ Mode Share	Low-Skill	High-Skill	Low-Skill	High-Skill	Low-Skill	High-Skill
Minibus	0	0	0.036	0.028	0.036	0.028
Car	0	0	-0.007	-0.015	-0.007	-0.015
%Δ Welfare	0.004	-0.01	0.13	0.04	0.15	0.07

Notes: Table summarizes key outcomes under implementation of each element of the social planner optimum, including the changes in the minibus and car mode shares by skill group as well as the percent change in group-level welfare, measured as equivalent variation, relative to the status quo. The first two columns show changes under optimal minibus supply b^*_{ij} alone, with commuter choice probabilities as in the status quo decentralized equilibrium. The third and fourth columns show changes under optimal commuter choices π^{g*}_{ijm} alone, and the last two columns under both simultaneously, i.e. the full social optimum.

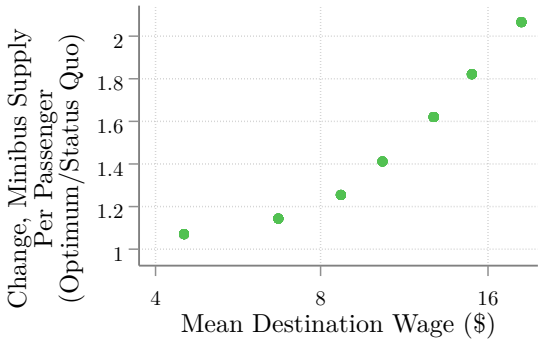
I begin with the optimal commuter choices. In the status quo, associations' limited form of market power from bargaining power $\beta > 0$ suppresses minibus demand. Meanwhile, Cape Town's relatively large-$\overline{\eta}$ minibuses leave substantial scope to reduce loading times and thus realize Market-Mohring increasing returns. In consequence, the planner increases passenger volumes on all routes, with a median increase of 15%. Figure 11a displays a binned scatterplot of changes in minibus passenger arrivals λ_{ij} versus average wages at a route's destination and shows that the planner would increase demand most strongly on routes to *lower*-wage destinations. These routes typically have lower baseline demand, longer waits, and therefore greater scope

FIGURE 11. SOCIAL PLANNER: DEMAND AND MINIBUS SUPPLY

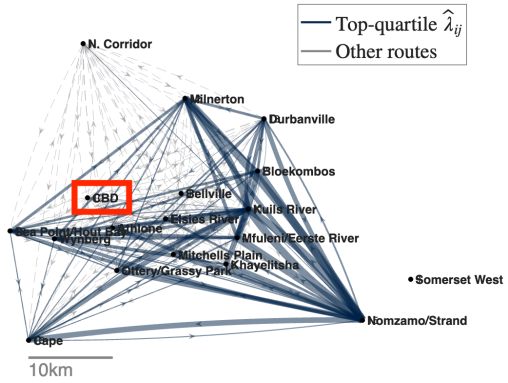
(A) OPTIMAL MINIBUS DEMAND VERSUS WAGES



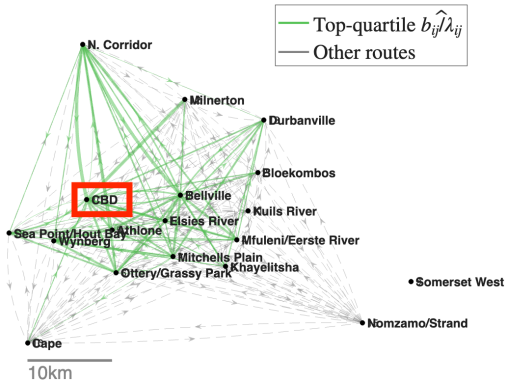
(B) OPTIMAL MINIBUS SUPPLY VERSUS WAGES



(C) LARGEST DEMAND INCREASES



(D) LARGEST SUPPLY INCREASES



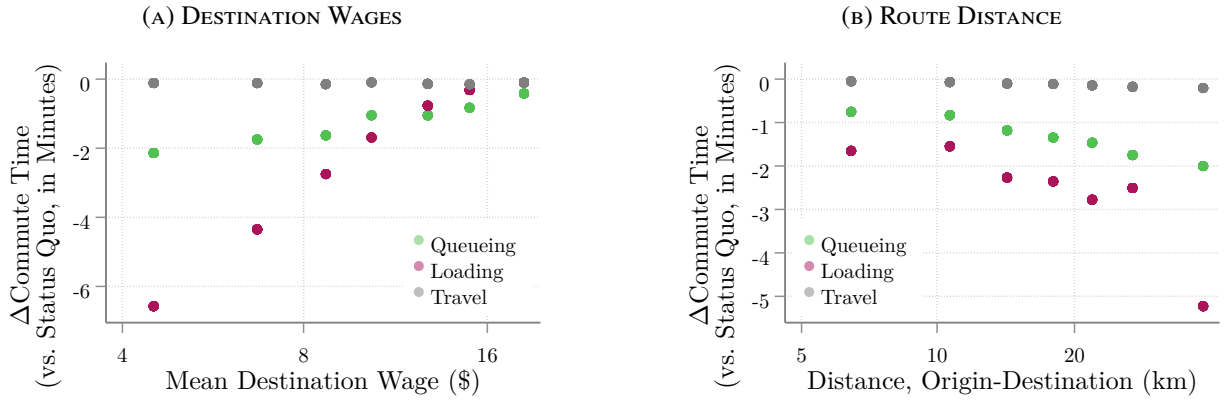
Notes: Panels (A) and (B) display, on the vertical axis, the social planner's optimal minibus passenger arrival rate λ^*_{ij} and optimal minibus supply per passenger, b^*_{ij}/λ^*_{ij}, respectively, in changes relative to the status-quo at the route level. The horizontal axis displays the average across skill groups, weighted by aggregate populations, of the corresponding work-location wage. Panels (C) and (D) map the 25% of routes with the largest proportionate increases in demand, λ_{ij}, and bus supply per passenger, b_{ij}/λ_{ij}, respectively, from the status quo to the social optimum. Line width indicates the magnitude by which the social optimum exceeds the status quo, and the CBD is outlined in red.

for Market-Mohring wait time reductions as buses fill faster. The map in Figure 11c shows that the largest demand increases occur on less-busy routes between far-flung suburbs, where association market power had previously limited demand and slowed bus loading.

Not only minibus demand but also supply turns out to fall short of the optimum. Nontrivial association bargaining power raises fares enough to suppress demand below its optimal scale, while $\beta < 1$ means associations still do not fully internalize the queue-reducing benefits of additional Bus Supply Insurance. In consequence, the planner increases minibus supply across the board, by a median of 46%. Unlike demand, the largest supply increases occur on the high-wage, high-demand routes most prone to queues. Spatially, in Figure 11d, an opposite set of routes again benefit: busy radial routes to the CBD require the largest proportionate increases in bus supply per passenger.

The planner's demand and supply choices leverage both Market-Mohring increasing returns and Bus Supply Insurance to reduce minibus commute times. Figures 12a-12b plot route-level changes in queueing Q_{ij},

FIGURE 12. SOCIAL PLANNER: CHANGE IN MINIBUS COMMUTE TIMES, VERSUS. . .



Notes: Figure characterizes the planner's optimum. Panel (A) displays, on the horizontal axis, the average across skill groups, weighted by aggregate populations, of the wage at a route's destination, and on the vertical axis, a binned scatterplot of route-level raw changes in queueing, loading, and travel times, Q_{ij}, L_{ij}, and T_{ij}, from the decentralized equilibrium status quo to the social optimum. Panel (B) instead displays a binned scatterplot of changes in each commute time component versus the straight-line distance from a minibus route's origin to destination.

loading L_{ij}, and on-road travel times T_{ij} against destination wages and route distance, respectively. Total queueing plus loading time declines, especially on low-wage and long routes.[42] These low-demand routes initially suffered the longest waits, both in the queue for a bus to arrive and on-board the bus for it to fill up, so they gain the most when the planner jointly scales up demand and bus supply. Table 5 also highlights a shift from cars to *shared* minibuses, which de-congests roads and, as also evident in Figures 12a-12b, modestly decreases on-road travel times. On the whole, however, the planner achieves only modest wait time savings — a first hint that Cape Town's privatized shared transit market performs better than its reputation. Indeed, the status quo lies close to the social optimum in welfare terms: low-skill commuters gain 0.15% in equivalent variation, and the high-skill gain 0.07%.[43]

Decomposition: Market Mohring Versus Bus Supply Insurance

To separate the Market Mohring and Bus Supply Insurance Effects, I simulate two counterfactuals that impose, one at a time, (i) optimal minibus supply b^*_{ij} and (ii) optimal commuter choice probabilities, π^{g*}_{ijm}. The first isolates the Bus Supply Insurance Effect, while the second identifies the Market Mohring Effect. The first two sets of columns in Table 5 show that optimization of bus supply alone yields negligible welfare gains, whereas the optimal commuter choices alone deliver 87% and 57% of the low- and high-skill gains from the full optimum. Thus, the Market Mohring Effect drives the lion's share of the relatively modest efficiency gains.

[42] Underlying these loading times is of course the assumption that buses depart full, a behavior which, in reality, may reflect unmodeled driver incentives, particularly for the subset of drivers paid on commission. Nonetheless, since the planner would increase passenger inflows, thereby decreasing loading times, the choice to depart full would seem unlikely to change under the optimal demand and supply.

[43] Note that association profits in the status quo are only 0.62% as high as consumer surplus. As a result, the gain in surplus from the social optimum equals 7.7% of baseline association profits, potentially leaving substantial scope for any future policy implementing the social optimum to split surplus between associations and commuters.

Nonetheless, the combined gains exceed those from optimal commuter choices alone, which underscores the complementarity between these two sources of increasing returns.

Efficiency and Market Power

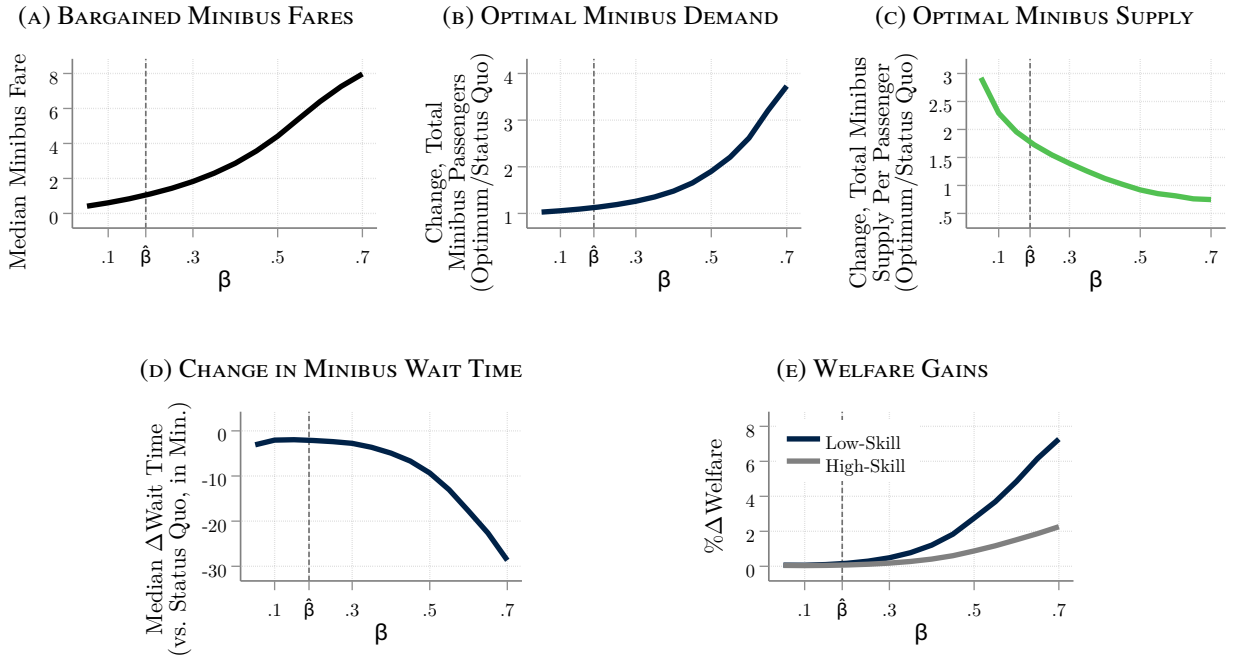
I now investigate when this key finding generalizes: which types of markets and cities should (not) expect efficient privatized provision? To explore alternative market conditions, I vary the key parameter governing equilibrium minibus demand and supply: association bargaining power, β. For each β, I recompute the decentralized equilibrium and compare it to the social optimum, which itself is independent of bargaining power. Figure 13a shows that greater market power in the form of higher β raises median fares. Accordingly, total minibus demand, i.e. the sum of route-level λ_{ij}, falls, and the planner's optimal demand increase in Figure 13b grows with β. Cape Town's relatively low bargaining power, at $\hat{\beta} = .19$, therefore induces a near-efficient demand scale that largely internalizes the Market Mohring Effect. At the same time, higher β facilitates internalization of Bus Supply Insurance benefits: as discussed in Section IV, it encourages associations to provide greater bus supply per passenger, so the planner's required supply increase in Figure 13c shrinks. Nonetheless, buses arrive quickly enough even at low supply levels that the Market Mohring Effect dominates. In Figure 13d, higher association bargaining power generates increasingly large median gaps in total minibus wait time $Q_{ij} + L_{ij}$ between the social optimum and the decentralized equilibrium. Accordingly, as shown in Figure 13e, the welfare gains from moving to the social optimum rise sharply with β. Thus, relatively low minibus-association market power underlies the efficiency of Cape Town's privatized shared transit market.[44]

Robustness

In Online Appendix E.3, I test robustness to five modifications: (i) higher queueing efficiency μ; (ii) higher bus arrival efficiency ρ; (iii) nested logit demand; (iv) agglomeration spillovers from local employment to wages; and (v) lower "non-rush-hour" passenger inflows. Across all specifications, the gains from (re-)optimization of demand and supply remain qualitatively similar. Agglomeration most strongly increases the gains because the planner can concentrate workers in ex-ante high-wage locations and further raise earnings. Furthermore, in Online Appendix E.4, I examine alternative market structures and optimization margins. Allowing the city government to bargain over bus supply, in addition to fares, moves the decentralized equilibrium even closer to the optimum. By contrast, forcing all buses to depart "early," before reaching capacity, yields no additional welfare gains for the social planner.

[44]Taking $\beta \to 0$ as approximation of government-set pricing and $\beta \to 1$ as approximation of deregulated pricing, the results in Figure 13e suggest that a greater public role in minibus pricing, possibly as part of fully-public provision, might improve efficiency in Cape Town. Indeed, $\beta \to 1$ is an extreme case of the exclusionary effects of fee pricing of public goods often studied in contexts such as education and health (Fabre and Straub (2023)). The fact that $\beta \to 0$ induces closer-to-efficient provision echoes the possibility of constrained-efficient private provision of public goods in Norman (2004), though his analysis applies only to public goods which, unlike transport, have zero marginal cost. Nevertheless, this rough thought experiment ignores the costs and feasibility of such alternative market structures and thus should not be interpreted as a policy counterfactual.

Figure 13. Optimum vs. Decentralized Equilibrium under Different Association Bargaining Power



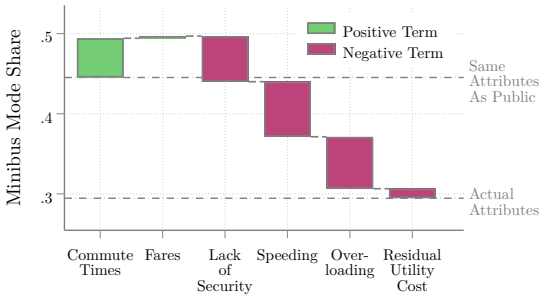
Notes: Figure displays different outcomes related to the decentralized equilibrium and the social optimum under different values of association bargaining power, β, displayed on the horizontal axis, with the estimated $\hat{\beta}$ indicated. Panel (A) displays the median bargained equilibrium minibus fare. Panel (B) displays the total minibus demand, calculated as the sum of passenger arrival rates λ_{ij} across routes, in the social optimum relative to the decentralized equilibrium. Panel (C) displays the total minibus supply per passenger in the optimum relative to the decentralized equilibrium, calculated in each case as the sum across routes of bus supply b_{ij} divided by the total minibus demand in Panel (B). Panel (D) displays the median change, from equilibrium to the social optimum, in the total minibus wait time $Q_{ij} + L_{ij}$ across routes. Finally, Panel (E) displays the welfare gains from the social optimum relative to the decentralized equilibrium, separately for low- and high-skill commuters. All other parameters remain as estimated.

A Decomposition of Minibus Demand

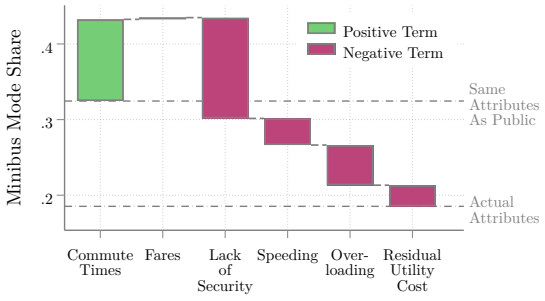
Finally, I ask why, despite its relative efficiency, Cape Town's minibus network still attracts limited demand from the aging public bus and train system, as evident in Figure 3a. Using Equation (8) and stated-preference valuations, I decompose the equilibrium aggregate minibus mode share into contributions from minibuses' relative advantages and disadvantages vis-à-vis public buses and trains. I consider commute times, fares, and different dimensions of quality. In the latter case, consistent with anecdotal evidence, I assume that minibuses currently lack security guards, exceed speed limits, and carry more passengers than seats. In Figures 14a-14b, I report the results of this decomposition by skill group. For each group, I set minibus commute times (wait plus travel time), fares, and utility costs equal to the public bus/train equivalents. The upper dotted line indicates the implied minibus mode share under attribute parity. From left to right, each bar then shows the incremental change in minibus mode share as I reset each minibus attribute to its observed value. To quantify the effects of the lack of security, of speeding, and of overloading, I add the corresponding estimated stated-preference valuations $|\xi^g|$ to the minibus utility cost κ_M^g. The final column restores the full estimated minibus utility cost and therefore recovers the model-predicted equilibrium minibus mode share, shown by

FIGURE 14. DECOMPOSITION: WHY DO(N'T) COMMUTERS USE MINIBUSES?

(A) LOW-SKILL



(B) HIGH-SKILL



Notes: Figure displays a decomposition of the aggregate minibus mode share, separately in Panel (A) for low-skill and in Panel (B) for high-skill commuters. In each plot, the upper dotted line indicates the aggregate minibus mode share with minibus commute times (wait plus travel $Q_{ij} + L_{ij} + T_{ij}$) and fares τ_{ij} on each origin-destination pair set equal to their public bus/train equivalents as well as a minibus utility cost κ_M^g set equal to the public bus/train utility cost. Each bar displays the (additional) change in minibus mode share when I set, in order from left to right, each component of minibus commute utility equal to its actual value(s). In order to quantify the effects of a lack of security, speeding, and overloading, I add the corresponding estimated stated preference effects ξ^g to the minibus utility cost. The last column then sets the minibus utility cost equal to its true value, generating the actual equilibrium minibus mode share indicated by the lower dotted line.

the lower dotted line.

The decomposition shows a clear commute-time versus quality tradeoff. If minibuses had the same commute times, fares, and an equally-low utility cost as public buses and trains, they would attain an almost 15 percentage-point higher mode share—the difference between the upper and lower dotted lines in Figures 14a-14b. The first bar in each figure, in turn, shows that, if minibuses benefit from their observed commute-time advantage but otherwise remain identical to the public modes, they would capture a 5–11 percentage-point larger mode share. Next, fares are similar across the two modes and thus have little impact on mode shares. In the next three bars, I then increase minibus utility costs using stated-preference valuations for station security, speed-limit compliance, and no overloading to reflect the fact that minibuses largely lack these attributes in the status quo. Each adjustment sharply lowers predicted minibus use; for instance, in the third bar of Figure 14b, accounting for the lack of station security guards reduces the high-skill minibus mode share by 13 percentage points.[45] Overall, then, minibuses in Cape Town deliver short commute times at the cost of crime risk, driving safety, and comfort.

[45]While this exercise suggests that reductions in speeding would increase the minibus mode share, these calculations ignore the likely ensuing increases in travel times. However, my demand model implies that minibus travel times would need to almost double for high-skill commuters and more than triple for low-skill commuters to fully cancel out the positive effect of eliminating speeding on the minibus mode share implied by Figures 14a-14b.

VIII. Conclusion

In this paper, I build the first model of the privatized shared transit sector that dominates many developing-country cities. Unlike existing frameworks that treat commute costs as fixed, my model predicts how commuter and transit provider choices give rise to equilibrium wait times, travel times, and fares. The wait time inherent to any shared transit arises from a queueing model that captures two sources of increasing returns: the Market Mohring Effect, whereby wait times decline with passenger demand, and the Bus Supply Insurance Effect, whereby bus supply prevents the growth of passenger queues. I collect new data on passenger queues and bus arrivals in Cape Town to directly estimate key queueing parameters. Furthermore, I introduce the stated preference approach to the urban literature to identify commuter demand primitives. Finally, I compare the existing decentralized minibus equilibrium with the social planner's optimum. Minibus associations' limited market power weakens incentives to improve service quality through additional bus supply but also keeps demand near its efficient scale and thereby reduces wait times close to optimally. The case of Cape Town therefore demonstrates that even seemingly-chaotic privatized shared transit systems can efficiently serve sprawling developing-country urban areas when cities get market power right.

Appendix

A. Theory

A.1 Queue Simulation

In this section, I discuss how I simulate my queueing model out of steady state for each route ij. For readability, I suppress the ij-subscript throughout this section. First, I derive the expected queueing time conditional on queue length and whether a bus is present in the station when a passenger arrives. Suppose that a passenger arrives to find n passengers in the queue for a given route and no bus present in the loading lane. Then, the expected queuing time Q until the passenger finishes boarding the bus satisfies

$$E\left(Q|n, \text{no bus present}\right) \equiv \frac{1}{\lambda^B} + \frac{n+1}{\mu p^b}. \tag{A.1}$$

The first term in (A.1) captures the wait for a bus to arrive. The second term equals the expected boarding time per passenger, $1/\mu$, times the number of passengers ahead of and including the newly-arrived passenger, $n+1$, divided by the unconditional probability, call it p^b, that a bus is present in the loading area.[46] This probability that a bus is present equals the ratio of expected loading time L, to the sum of loading time and the gap $1/\lambda^B$ between buses, such that

$$p^b \equiv \frac{L}{\frac{1}{\lambda^B} + L}. \tag{A.2}$$

[46] Intuitively, imagine a bus is present only half the time; in that case, the time elapsed, even after an initial bus arrives, until the newly-arrived passenger boards will be twice as long as if a bus were continuously present and able to board passengers.

In turn, if a passenger arrives to find a bus present and loading, expected queueing time no longer needs to account for the initial bus gap and satisfies $E\left(Q|n, \text{bus present}\right) \equiv (n+1)\,/\left(\mu p^b\right)$.

Second, I consider how the queue for a given route evolves in continuous time over some instants indexed by t. Denote the (joint) probability that, at instant t, n passengers wait in the queue and a bus is present as $p_n^b(t)$ and the probability that n passengers wait in the queue and *no* bus is present as $q_n(t)$. As in Brancaccio et al. (2024), each of these probabilities evolves according to Kolmogorov Forward Equations, as follows. For the pdf of queue lengths with a bus present:

$$\frac{d}{dt}p_n^b(t) = \begin{cases} -\left(\lambda + \mu^B(t)\right) p_0^b(t) + \lambda^B q_0(t) + \mu p_1^b(t), & \text{if } n = 0 \\ -\left(\lambda + \mu + \mu^B(t)\right) p_n^b(t) + \lambda^B q_n(t) + \mu p_{n+1}^b(t) + \lambda p_{n-1}^b(t), & \text{if } n > 0 \end{cases} \tag{A.3}$$

For interpretation, consider the case of no queue but a bus present and loading, where $n = 0$. Over time, this probability decreases by an "outflow" in the first term equal to the probability of that state times the rate λ at which another passenger arrives, taking the queue into state $(n = 1, \text{bus present})$, plus the (time-varying) rate $\mu^B(t)$ at which the bus departs, taking the queue into state $(n = 0, \text{no bus})$. The probability $p_0^b(t)$ increases by the probability of the state $(n = 0, \text{no bus})$ times the rate λ^B at which buses arrive plus the probability of a queue of 1 with a bus present times the rate μ at which that passenger boards the bus. The evolution of $p_n^b(t)$ for $n > 0$ must additionally account for an outflow of passengers boarding that bus at rate μ as well as an increase over time in this probability by the probability of a one-passenger-shorter queue times the rate at which an additional passenger arrives.

The evolution of the probability of the states $(n, \text{no bus})$ follows a similar logic,

$$\frac{d}{dt}q_n(t) = \begin{cases} -\left(\lambda + \lambda^B\right) q_0(t) + \mu^B(t) p_0^b(t), & \text{if } n = 0 \\ -\left(\lambda + \lambda^B\right) q_n(t) + \mu^B(t) p_n^b(t) + \lambda q_{n-1}(t), & \text{if } n > 0 \end{cases} \tag{A.4}$$

whereby this probability, for the general n case, decreases to an extent determined by the rate at which new passengers arrive and the rate λ^B at which a bus arrives. $q_n(t)$ increases by an inflow equal to the probability of n passengers and a bus loading times the rate $\mu^B(t)$ at which that bus departs plus the likelihood of a one-person shorter queue and no bus loading, $q_{n-1}(t)$ times the passenger arrival rate.

I then consider a discrete grid of $T = 300$ instants $\{t_1, \ldots, t_T\}$, each 12 seconds apart and indexed by h, such that my total simulation captures the queue's evolution over 1 hour. I take λ^B, μ, and λ as given and constant over the hour. I initialize the queue at $q_0(0) \approx 1$ and then, again following Brancaccio et al. (2024), apply a first-order approximation to characterize the pdfs of queue lengths at each instant:

$$p_n^b(t_h) = p_n^b(t_{h-1}) + 0.2 \times \frac{d}{dt}p_n^b(t_{h-1}), \text{ and}$$
$$q_n(t_h) = q_n(t_{h-1}) + 0.2 \times \frac{d}{dt}q_n(t_{h-1}).$$

Following the equations given in the main text, for each instant h, I then calculate $\mu^B(t_h) = \frac{\mu}{\eta}\frac{\sum_{n>0} p_n^b(t_h)}{\sum_n p_n^b(t_h)}$, and, from Equation (A.2), I calculate $p^b(t_h) = \frac{1/\mu^B(t_h)}{1/\mu^B(t_h)+1/\lambda^B}$. Finally, I calculate expected queueing and loading times as

$$\begin{aligned} Q &= \frac{1}{T}\sum_h \sum_n \left[p_n^b(t_h)E(Q(t_h)|n, \text{bus present}) + q_n(t_h)E(Q(t_h)|n, \text{no bus present})\right] \\ &= \frac{1}{T}\sum_h \sum_n \left[p_n^b(t_h)\left(\frac{n+1}{\mu p^b(t_h)}\right) + q_n(t_h)\left(\frac{1}{\lambda^B} + \frac{n+1}{\mu p^b(t_h)}\right)\right], \end{aligned}$$

where the second equality substitutes in the expectation in Equation (A.1) as well as the corresponding equation for the case where a bus is present, and

$$L = \frac{1}{T}\sum_h \frac{1}{\mu^B(t_h)}.$$

By substituting the given λ^B, the (expected) L obtained through this simulation, and the route's on-road travel time T into the Poisson relation $\frac{1}{\lambda^B} \equiv \frac{1}{\rho}\log\left\{\exp\left[\rho\left(\frac{2T}{b} - L\right)\right] + 1\right\}$, I obtain the necessary bus supply b.[47]

A.2 Steady-State Queue Derivations

In this section, I derive key probabilities and queueing times for my queueing model in steady-state. Before proceeding to the derivations, I establish equivalence between quantities in my framework and the one-sided M/M/1 queue with service interruptions in Avi-Itzhak and Naor (1963). The notation is complicated by the fact that they, in some cases, use the same variables to denote different quantities. For clarity, in the remainder of this paragraph, I write equivalencies with their notation on the left-hand side and mine on the right-hand side. They denote the arrival rate of customers to the queue as $\xi \equiv \lambda_{ij}$. Their fraction b of time that the "service station" is busy is equivalent to the fraction of time during which passengers are boarding a loading bus, $b \equiv \frac{\lambda_{ij}}{\mu}$, while the parameter λ of the exponential distribution of durations of uninterrupted availability is equivalent to the rate at which loading buses depart, $\lambda \equiv \mu_{ij}^B$. Their expected time until a broken service station is repaired, $E(t_r)$, is equivalent to the expected time between one bus departure and the instant when the next bus begins loading, which, given Poisson bus arrivals, equals the inverse of the corresponding arrival rate, or $E(t_r) \equiv \frac{1}{\lambda_{ij}^B}$. Finally, Avi-Itzhak and Naor (1963)'s probability p_0 that the service station is in order equals the probability that a bus is present in the station, which, as in the preceding section, I henceforth denote by p_{ij}^b.

First, I derive key probabilities and expected loading times. I start with the unconditional probability p_{ij}^b that a bus is present in the station. To obtain this probability, I first discuss a key flow balance condition. In steady-state, the average rate at which passengers depart on buses equals the probability that a bus is present times the rate μ at which passengers board, adjusted by the fraction of time a queue exists conditional on a bus

[47] For computational reasons, I set a maximum queue length N, sufficiently high such that $p_N^b(t)$ and $q_N(t)$ are small, and also solve only for every 15th $p_n^b(t)$ and $q_n(t)$ for n above some threshold $\tilde{N}$. I adjust the Kolmogorov Forward Equations (A.3)-(A.4) accordingly.

being present, such that this outflow of passengers on buses equals $p^b_{ij}\mu p_{ij}$. In steady state, this outflow must equal the inflow of passengers, i.e. $p^b_{ij}\mu p_{ij} = \lambda_{ij}$. Then, the probability that a bus is present must satisfy

$$p^b_{ij} = \frac{1}{1 + \frac{\mu p_{ij}}{\overline{\eta}}\left(\frac{1}{\lambda^B_{ij}}\right)} = 1 - \frac{\lambda_{ij}}{\overline{\eta}\lambda^B_{ij}}, \tag{A.5}$$

where the first equality uses Equation (7) in Avi-Itzhak and Naor (1963) and the second, the aforementioned steady-state flow balance condition $\lambda_{ij} = p^b_{ij}\mu p_{ij}$.

I now derive the conditional probability p_{ij} of a nonzero queue, given that a bus is present in the station. Well-known results for M/M/1 queues (Hillier and Lieberman (2024)) imply that, in steady state, the queue "operates," i.e. a bus is present and passengers actively board, a fraction of the time given by the ratio of the rate λ_{ij} at which passengers arrive to the queue to the rate μ at which they depart the queue. The conditional probability p_{ij} of a nonzero queue, given a bus in the station, then, equals the ratio of the former joint probability of a passenger queue and a bus being present to the unconditional probability of a bus being present derived in the preceding paragraph, or

$$p_{ij} = \frac{\frac{\lambda_{ij}}{\mu}}{1 - \frac{\lambda_{ij}}{\overline{\eta}\lambda^B_{ij}}}. \tag{A.6}$$

Inserting this probability into Equation (1) in the main text yields steady-state loading times,

$$L_{ij} = \frac{\overline{\eta}}{\lambda_{ij}} - \frac{1}{\lambda^B_{ij}}. \tag{A.7}$$

Second, I derive expected queueing time. My expected queueing time, Q_{ij}, is equivalent to the expected wait time in the queue plus service time in Avi-Itzhak and Naor (1963), which they denote by θ_q, and their expected number of customers in the system, q, is equivalent to queue length, n_{ij}, in my model. The equivalent of their Equation (22) in my model is $Q_{ij} = n_{ij}/\lambda_{ij}$. Avi-Itzhak and Naor (1963) provide an expression for the expected queue length, $q \equiv n_{ij}$, in their Equation (33).[48] Note that their fraction of time the service station is out of service, p_1, satisfies $p_1 = 1 - p_0 \equiv 1 - p^b_{ij}$, and that Equation (A.5) implies that $\frac{1}{\lambda^B_{ij}} = \left(1 - p^b_{ij}\right)\frac{\overline{\eta}}{\lambda_{ij}}$. I substitute these equivalencies into their Equation (33). Then, I substitute this equivalent of their Equation (33) for my n_{ij} into $Q_{ij} = n_{ij}/\lambda_{ij}$. Finally, I obtain an expression for the expected queueing time in my model,

$$Q_{ij} = \frac{\mu^{-1} + \frac{\overline{\eta}p^b_{ij}\left(1-p^b_{ij}\right)^2}{\lambda_{ij}}}{p^b_{ij} - \frac{\lambda_{ij}}{\mu}} \approx \frac{1}{\mu p^b_{ij} - \lambda_{ij}} = \frac{1}{\mu\left(1 - \frac{\lambda_{ij}}{\overline{\eta}\lambda^B_{ij}}\right) - \lambda_{ij}}, \tag{A.8}$$

[48] This equation applies to the case of my model since I assume Poisson bus arrivals (at rate λ^B_{ij}) and passenger boarding (at rate μ), which means that the exponentially-distributed gaps between successive buses as well as boarding times have coefficients of variation equal to 1.

where the (second) approximation holds for routes with a bus loading most of the time, $p^b_{ij} \to 1$ such that $\left(1 - p^b_{ij}\right)^2 \to 0$, and the final equality again uses Equation (A.5) .

A.3 Efficiency Proofs

I first derive a conveniently rewritten form of the welfare function Ω.

Lemma A.1. *Welfare satisfies*

$$\begin{aligned}\Omega = \sum_{i,j,g} N^g \pi^g_{ijM} \left[\theta^g_i - \kappa^g_M - r\omega^g_j \left(Q_{ij} + L_{ij} + T_{ij}\right) + \omega^g_j - \nu \log \pi^g_{ijM} - \frac{\chi_{ij}}{\overline{\eta}}\right] \\ + \sum_{i,j,g} N^g \pi^g_{ijP} \left[\theta^g_i - \kappa^g_P - r\omega^g_j \left(H_{ij} + T_{ijP}\right) - \tau_{ijP} + \omega^g_j - \nu \log \pi^g_{ijP}\right] \\ + \sum_{i,j,g} N^g \pi^g_{ijA} \left[\theta^g_i - \kappa^g_A - r\omega^g_j T_{ij} - \tau_A + \omega^g_j - \nu \log \pi^g_{ijA}\right] - \overline{\omega} \sum_{i,j} b_{ij}. \quad \text{(A.9)}\end{aligned}$$

Proof. In (10), consider first the ex-ante expected utility of group-g commuters, $\overline{\Omega}^g$, given their optimal choices of home, work, and mode, subject to idiosyncratic Gumbel-distributed preference shocks.[49] Denoting total deterministic utility of alternative ijm by $\overline{U}^g_{ijm} \equiv \theta^g_i + U^g_{ijm} + \omega^g_j$, I rewrite expected utility as

$$\begin{aligned}\overline{\Omega}^g &\equiv E\left[\max_{i',j',m'} \left(\overline{U}^g_{i'j'm'} + \nu\varepsilon_{i'j'm'}\right)\right] \\ &= \sum_{i,j,m} \pi^g_{ijm} \left[\overline{U}^g_{ijm} + \nu E\left(\varepsilon_{ijm} | ijm \in \text{argmax}_{i',j',m'} \left(\overline{U}^g_{i'j'm'} + \nu\varepsilon_{i'j'm'}\right)\right)\right] \\ &= \sum_{i,j,m} \pi^g_{ijm} \left[\theta^g_i + U^g_{ijm} + \omega^g_j - \nu \log \pi^g_{ijm}\right]. \quad \text{(A.10)}\end{aligned}$$

The final equality uses a well-known result that the expected value of Gumbel preference shocks of agents who have optimally chosen a given alternative equals the negative of the corresponding log choice probability.[50] Next, I substitute (9) into association profits (4) and sum across routes to obtain $\Pi \equiv \sum_{i,j} \Pi_{ij}$. Finally, substituting each element into (10), using the mode-specific definitions of commute value U^g_{ijm}, and rearranging, I obtain the expression in (A.9). □

[49] Note that the social planner chooses choice probabilities directly and would then implement these choice probabilities with appropriately-set home, work, and commute-mode-specific transfers to commuters, in the spirit of Donald et al. (2025).

[50] To see this, note that

$$\begin{aligned}E\left[\varepsilon_{ijm} | ijm \in \text{argmax}_{i',j',m'} \left(\overline{U}^g_{i'j'm'} + \nu\varepsilon_{i'j'm'}\right)\right] &= \frac{1}{\nu}\left[E\left[\max_{i',j',m'} \left(\overline{U}^g_{i'j'm'} + \nu\varepsilon_{i'j'm'}\right)\right] - \overline{U}^g_{ijm}\right] \\ = \frac{1}{\nu}\left[\nu \log\left[\sum_{i',j',m'} \exp\left[\overline{U}^g_{i'j'm'}\right]^{1/\nu}\right] - \overline{U}^g_{ijm}\right] &= \log\left[\sum_{i',j',m'} \exp\left[\overline{U}^g_{i'j'm'}\right]^{1/\nu} / \exp\left(\overline{U}^g_{ijm}\right)^{1/\nu}\right] = -\log \pi^g_{ijm}\end{aligned}$$

where the first equality uses the well-known property of discrete choice models with Gumbel shocks whereby the conditional equals the unconditional expected utility, the second substitutes in for $\overline{\Omega}^g$, and the final uses the choice probabilities in Equation (8).

Proof of Proposition 1

Proof. First, I derive the conditions characterizing the social planner optimum. From the definition of optimality in the main text, the planner solves

$$\max_{\{\pi\}_{i,j,m,g}\{b\}_{i,j}} \Omega \text{ s.t. } Q_{ij} = Q\left(\lambda_{ij}, \lambda_{ij}^B\right),\ L_{ij} = L\left(\lambda_{ij}, \lambda_{ij}^B\right),\ \lambda_{ij}^B = \mathrm{g}\left(b_{ij}, L_{ij}, T_{ij}; \rho\right),$$
$$T_{ij} = \sum_{kk' \in s(i,j)} \bar{t}_{kk'} x_{kk'}^{\gamma}, \text{ and } \sum_{i,j,m} \pi_{ijm}^{g} = 1. \quad \text{(A.11)}$$

I substitute the $Q(\cdot,\cdot)$ and $L(\cdot,\cdot)$ which result from my numerical queue simulations as well as the expression for link-level vehicle volume, $x_{kk'} = \sum_{i',j':kk' \in s(i',j')} \sum_{g'} N^{g'} \left(\pi_{i'j'M}^{g'}/\overline{\eta} + \pi_{i'j'A}^{g'}\right)$, into the expression in Lemma A.1 to obtain welfare as $\Omega(\boldsymbol{b}, \boldsymbol{\pi})$. I can thus rewrite the planner's problem as $\max_{\boldsymbol{b},\boldsymbol{\pi}} \Omega(\boldsymbol{b}, \boldsymbol{\pi})$ s.t. $\sum_{i,j,m} \pi_{ijm}^{g} = 1$.[51] The planner's first-order condition for bus supply b_{ij} on each route reads

$$\frac{\partial \Omega}{\partial b_{ij}} = -\sum_{g} N^g \pi_{ijM}^{g*} r \omega_j^g \left(\frac{\partial Q_{ij}}{\partial \lambda_{ij}^B} + \frac{\partial L_{ij}}{\partial \lambda_{ij}^B}\right) \frac{\partial \lambda_{ij}^B}{\partial b_{ij}} - \overline{\omega} = 0. \quad \text{(A.12)}$$

The first-order conditions for optimal commuter choice probabilities π_{ijm}^{g*}, in turn, can be combined with the condition for a reference choice probability for each group g, π_{klP}^{g*}, to derive an optimal relative choice probability:

$$\begin{aligned}
\frac{\pi_{ijm}^{g*}}{\pi_{klP}^{g*}} = \exp\Bigg(\theta_i^g - \kappa_m^g + \omega_j^g + \mathbb{1}\{m = M\} \Bigg\{ & -r\omega_j^g \left(Q_{ij} + L_{ij} + T_{ij}\right) - \frac{\chi_{ij}}{\overline{\eta}} \\
& - \sum_{i',j',g'} N^{g'} \pi_{i'j'M}^{g'*} r \omega_{j'}^{g'} \Bigg[\mathbb{1}\{i',j' = i,j\} \left(\frac{\partial Q_{i'j'}}{\partial \lambda_{i'j'}} + \frac{\partial L_{i'j'}}{\partial \lambda_{i'j'}}\right) + \\
& \left(\frac{\partial Q_{i'j'}}{\partial \lambda_{i'j'}^B} + \frac{\partial L_{i'j'}}{\partial \lambda_{i'j'}^B}\right) \left(\mathbb{1}\{i',j' = i,j\} \left.\frac{\partial \lambda_{i'j'}^B}{\partial \lambda_{i'j'}}\right|_{T_{i'j'}} + \frac{\partial \lambda_{i'j'}^B}{\partial T_{i'j'}} \frac{\partial T_{i'j'}}{\partial \pi_{ijM}^g} \frac{1}{N^g}\right) \Bigg] \\
& - \sum_{i',j',g'} N^{g'} \left(\pi_{i'j'M}^{g'*} + \pi_{i'j'A}^{g'*}\right) r \omega_{j'}^{g'} \frac{\partial T_{i'j'}}{\partial \pi_{ijM}^g} \frac{1}{N^g} \Bigg\} \\
+ \mathbb{1}\{m = A\} \Bigg[& -r\omega_j^g T_{ij} - \tau_A - \sum_{i',j',g'} N^{g'} \pi_{i'j'M}^{g'*} r \omega_{j'}^{g'} \left(\frac{\partial Q_{i'j'}}{\partial \lambda_{i'j'}^B} + \frac{\partial L_{i'j'}}{\partial \lambda_{i'j'}^B}\right) \left(\frac{\partial \lambda_{i'j'}^B}{\partial T_{i'j'}} \frac{\partial T_{i'j'}}{\partial \pi_{ijA}^g} \frac{1}{N^g}\right) \\
& - \sum_{i',j',g'} N^{g'} \left(\pi_{i'j'M}^{g'*} + \pi_{i'j'A}^{g'*}\right) r \omega_{j'}^{g'} \frac{\partial T_{i'j'}}{\partial \pi_{ijA}^g} \frac{1}{N^g} \Bigg] \\
+ \mathbb{1}\{m = P\} \Big[& -r\omega_j^g \left(H_{ij} + T_{ijP}\right) - \tau_{ijP} \Big] - \theta_k^g + \kappa_P^g + r\omega_l^g \left(H_{kl} + T_{klP}\right) + \tau_{klP} - \omega_l^g \Bigg)^{1/\nu}. \quad \text{(A.13)}
\end{aligned}$$

[51] This welfare function includes the expectation of the idiosyncratic shocks, under the assumption that the planner induces the optimal share of commuters of each skill group to choose each home, work, and mode tuple through appropriately-set transfers, as in Donald et al. (2025).

Note that I calculate $\frac{\partial Q_{ij}}{\partial \lambda^B_{ij}}, \frac{\partial L_{ij}}{\partial \lambda^B_{ij}}, \frac{\partial Q_{ij}}{\partial \lambda_{ij}}$, and $\frac{\partial L_{ij}}{\partial \lambda_{ij}}$ numerically. I calculate $\left.\frac{\partial \lambda^B_{i'j'}}{\partial \lambda_{i'j'}}\right|_{T_{i'j'}}$, $\frac{\partial \lambda^B_{i'j'}}{\partial T_{i'j'}}$, and, from the bus supply first-order condition, $\frac{\partial \lambda^B_{ij}}{\partial b_{ij}}$ using the implicit function theorem, and I calculate $\frac{\partial T_{i'j'}}{\partial \pi^g_{ijM}}$ as well as $\frac{\partial T_{i'j'}}{\partial \pi^g_{ijA}}$ using the chain rule and the expressions introduced in the main text. Equations (A.12), (A.13), and the adding-up constraints in (A.11) together define the *social planner's allocation* and can be solved for the I^2 optimal bus supply levels b^*_{ij} as well as the $G \cdot I^2 \cdot 3$ optimal commuter choice probabilities π^{g*}_{ijm}.

Second, I derive conditions under which the associations and commuters internalize the potential sources of inefficiency, as defined in the main text. Starting with bus supply, equilibrium supply b_{ij} from (5) coincides with the optimal entry b^*_{ij} pinned down by the planner's bus entry condition (A.12) if and only if

$$\left(\tau_{ij} - \frac{\chi_{ij}}{\overline{\eta}}\right) \sum_g N^g \frac{\partial \pi^g_{ijM}}{\partial b_{ij}} = -\sum_g N^g \pi^{g*}_{ijM} r\omega^g_j \left(\frac{\partial Q_{ij}}{\partial \lambda^B_{ij}} + \frac{\partial L_{ij}}{\partial \lambda^B_{ij}}\right) \frac{\partial \lambda^B_{ij}}{\partial b_{ij}}. \tag{A.14}$$

Rearranging yields (12) in the main text.

Next, turning to demand, I must consider the planner's first-order conditions with respect to commuter choice probabilities π^g_{ijm}, holding travel times $\boldsymbol{T}$ fixed, to correspond to the definition of the Market Mohring Effect. These conditions can be combined with the condition for a reference choice probability for each group g, $\pi^{g*|T}_{klP}$, to derive an optimal relative choice probability, holding $\boldsymbol{T}$ fixed:

$$\begin{aligned}\frac{\pi^{g*|T}_{ijm}}{\pi^{g*|T}_{klP}} = \exp\Bigg(\theta^g_i - \kappa^g_m + \omega^g_j + \mathbb{1}\{m = M\}\Bigg\{-r\omega^g_j\left(Q_{ij} + L_{ij} + T_{ij}\right) - \frac{\chi_{ij}}{\overline{\eta}} \\ - \sum_{g'} N^{g'} \pi^{g'*|T}_{ijM} r\omega^{g'}_j \left[\frac{\partial Q_{ij}}{\partial \lambda_{ij}} + \frac{\partial L_{ij}}{\partial \lambda_{ij}} + \left(\frac{\partial Q_{ij}}{\partial \lambda^B_{ij}} + \frac{\partial L_{ij}}{\partial \lambda^B_{ij}}\right) \left.\frac{\partial \lambda^B_{ij}}{\partial \lambda_{ij}}\right|_{T_{ij}}\right]\Bigg\} \\ + \mathbb{1}\{m = A\}\left[-r\omega^g_j T_{ij} - \tau_A\right] \\ + \mathbb{1}\{m = P\}\left[-r\omega^g_j\left(H_{ij} + T_{ijP}\right) - \tau_{ijP}\right] - \theta^g_k + \kappa^g_P + r\omega^g_l\left(H_{kl} + T_{klP}\right) + \tau_{klP} - \omega^g_l\Bigg)^{1/\nu}.\end{aligned} \tag{A.15}$$

I substitute commute values U^g_{ijm} into (8); it is then immediate that equilibrium (relative) choice probabilities, π^g_{ijm}/π^g_{klP}, equal those chosen by the planner, holding travel times fixed, in (A.15) if and only if bargained minibus fares compensate, for each route and skill group, per-passenger operations costs plus the marginal value of the Market Mohring Effect, as in (11).

In a similar vein, I take the planner's first-order conditions with respect to commuter choice probabilities π^g_{ijm}, holding fixed the minibus passenger arrival rates $\boldsymbol{\lambda}$ which enter the queueing time and loading time functions, $Q(\cdot,\cdot)$ and $L(\cdot,\cdot)$, to correspond to the definition of the Road Congestion Externality. I again derive an optimal relative choice probability for each origin-by-destination-by-mode-by-skill, holding $\boldsymbol{\lambda}$ fixed:

$$\frac{\pi^{g*|\lambda}_{ijm}}{\pi^{g*|\lambda}_{klP}} = \exp\left(\theta^g_i - \kappa^g_m + \omega^g_j + \mathbb{1}\{m = M\}\left\{-r\omega^g_j\left(Q_{ij} + L_{ij} + T_{ij}\right) - \frac{\chi_{ij}}{\overline{\eta}}\right.\right.$$
$$\left. - \sum_{i',j',g'} N^{g'}\pi^{g'*|\lambda}_{i'j'M} r\omega^{g'}_{j'}\left(\frac{\partial Q_{i'j'}}{\partial\lambda^B_{i'j'}} + \frac{\partial L_{i'j'}}{\partial\lambda^B_{i'j'}}\right)\frac{\partial\lambda^B_{i'j'}}{\partial T_{i'j'}}\frac{\partial T_{i'j'}}{\partial\pi^g_{ijM}}\frac{1}{N^g}\right.$$
$$\left.\left. - \sum_{i',j',g'} N^{g'}\left(\pi^{g'*|\lambda}_{i'j'M} + \pi^{g'*|\lambda}_{i'j'A}\right) r\omega^{g'}_{j'}\frac{\partial T_{i'j'}}{\partial\pi^g_{ijM}}\frac{1}{N^g}\right\}\right.$$
$$+ \mathbb{1}\{m = A\}\left[-r\omega^g_j T_{ij} - \tau_A - \sum_{i',j',g'} N^{g'}\pi^{g'*|\lambda}_{i'j'M} r\omega^{g'}_{j'}\left(\frac{\partial Q_{i'j'}}{\partial\lambda^B_{i'j'}} + \frac{\partial L_{i'j'}}{\partial\lambda^B_{i'j'}}\right)\left(\frac{\partial\lambda^B_{i'j'}}{\partial T_{i'j'}}\frac{\partial T_{i'j'}}{\partial\pi^g_{ijA}}\frac{1}{N^g}\right)\right.$$
$$\left. - \sum_{i',j',g'} N^{g'}\left(\pi^{g'*|\lambda}_{i'j'M} + \pi^{g'*|\lambda}_{i'j'A}\right) r\omega^{g'}_{j'}\frac{\partial T_{i'j'}}{\partial\pi^g_{ijA}}\frac{1}{N^g}\right]$$
$$+\mathbb{1}\{m = P\}\left[-r\omega^g_j\left(H_{ij} + T_{ijP}\right) - \tau_{ijP}\right] - \theta^g_k + \kappa^g_P + r\omega^g_l\left(H_{kl} + T_{klP}\right) + \tau_{klP} - \omega^g_l\Big)^{1/\nu}. \quad \text{(A.16)}$$

Comparison with the equilibrium relative choice probabilities based on (8) makes it immediate that car choice probabilities never equal these constant-λ_{ij} optimal probabilities, i.e. $\pi^g_{ijA} \neq \pi^{g*|\lambda}_{ijA}$. Note that minibus commuters could in theory internalize their road congestion spillovers in another knife-edge case, $\tau_{ij} = \frac{\chi_{ij}}{\overline{\eta}} + \sum_{i',j',g'} N^{g'}\pi^{g'*|\lambda}_{i'j'M} r\omega^{g'}_{j'}\left(\frac{\partial Q_{i'j'}}{\partial\lambda^B_{i'j'}} + \frac{\partial L_{i'j'}}{\partial\lambda^B_{i'j'}}\right)\frac{\partial\lambda^B_{i'j'}}{\partial T_{i'j'}}\frac{\partial T_{i'j'}}{\partial\pi^g_{ijM}}\frac{1}{N^g} + \sum_{i',j',g'} N^{g'}\left(\pi^{g'*|\lambda}_{i'j'M} + \pi^{g'*|\lambda}_{i'j'A}\right) r\omega^{g'}_{j'}\frac{\partial T_{i'j'}}{\partial\pi^g_{ijM}}\frac{1}{N^g}$. However, commuters will never universally internalize the Road Congestion Externality. □

References

Abenoza, Roberto F., Oded Cats, and Yusak O. Susilo. 2019. "How does travel satisfaction sum up? An exploratory analysis in decomposing the door-to-door experience for multimodal trips." *Transportation* 46:1615–1642.

Ahlfeldt, Gabriel M., Stephen J. Redding, Daniel M. Sturm, and Nikolaus Wolf. 2015. "The Economics of Density: Evidence from the Berlin Wall." *Econometrica* 83 (6): 2127–2189.

Almagro, Milena, Felipe Barbieri, Juan Camilo Castillo, Nathaniel Hickok, and Tobias Salz. 2024. "Optimal Urban Transportation Policy: Evidence from Chicago."

Anderson, Brile, Rafael Prieto-Curiel, and Jorge Patino. 2023. "City Shapes and Climate Change in Africa." OECD/SWAC West African Papers No. 40.

Angel, Shlomo, Alex Blei, Patrick Lamson-Hall, Nicolas Galarza Sanchez, Pritha Gopalan, Achilles Kallergis, Daniel L. Civco, et al. 2016. *Atlas of Urban Expansion.* Technical report. NYU Urban Expansion Program, UN-Habitat, and the Lincoln Institute of Land Policy. http://www.atlasofurbanexpansion.org.

Ansari Esfeh, Mohammad, S. C. Wirasinghe, Saeid Saidi, and Lina Kattan. 2021. "Waiting time and headway modelling for urban transit systems – a critical review and proposed approach." *Transport Reviews* 41 (2): 141–163.

Antrobus, Lauren, and Andrew Kerr. 2019. "The labour market for minibus taxi drivers in South Africa." SALDRU Working Paper No. 250.

Avi-Itzhak, B., and P. Naor. 1963. "Some Queueing Problems with the Service Station Subject to Breakdown." *Operations Research* 11 (3): 303–320.

Balboni, Clare, Gharad Bryan, Melanie Morten, and Bilal Siddiqi. 2020. "Transportation, Gentrification, and Urban Mobility: The Inequality Effects of Place-Based Policies."

Banares-Sanchez, Ignacio, and Yoshiki Wiskamp. 2025. "Trading Trash on Tricycles."

Bandiera, Oriana, Michael Callen, Katherine E. Casey, Eliana La Ferrara, Camille Landais, Matthieu Teachout, and Shahid Vaziralli. 2024. *State effectiveness.* Technical report. IGC Evidence Paper. https://www.theigc.org/sites/default/files/2024-07/IGC%20State%20effectiveness%20Evidence%20paper%20July%202024.pdf.

Barwick, Panle Jia, Shanjun Li, Andrew R. Waxman, Jing Wu, and Tianli Xia. 2022. "Efficiency and Equity Impacts of Urban Transportation Policies with Equilibrium Sorting." *Revise & Resubmit, American Economic Review*.

Ben-Akiva, Moshe, Daniel McFadden, and Kenneth Train. 2019. "Foundations of Stated Preference Elicitation: Consumer Behavior and Choice-based Conjoint Analysis." *Foundations and Trends in Econometrics* 10 (1–2): 1–144.

Besley, Timothy, and Maitreesh Ghatak. 2001. "Government Versus Private Ownership of Public Goods." *The Quarterly Journal of Economics* 116 (4): 1343–1372.

Björkegren, Daniel, Alice Duhaut, Geetika Nagpal, and Nick Tsivanidis. 2025. "Public and Private Transit: Evidence from Lagos."

Borghorst, Malte, Ismir Mulalic, and Jos van Ommeren. 2021. "Commuting, children and the gender wage gap."

Borker, Girija. 2021. "Safety First: Perceived Risk of Street Harassment and Educational Choices of Women." World Bank Policy Research Working Paper 9731, *Revise & Resubmit, American Economic Review*.

Brancaccio, Giulia, Myrto Kalouptsidi, and Theodore Papageorgiou. 2024. "Investment in Infrastructure and Trade: The Case of Ports."

Bryan, Gharad, Edward Glaeser, and Nick Tsivanidis. 2020. "Cities in the Developing World." *Annual Review of Economics* 12:21.1–21.25.

Bryan, Gharad, and Melanie Morten. 2019. "The Aggregate Productivity Effects of Internal Migration: Evidence from Indonesia." *Journal of Political Economy* 127 (5): 2229–2268.

Buchholz, Nicholas, Laura Doval, Jakub Kastl, Filip Matejka, and Tobias Salz. 2024. "Personalized Pricing and the Value of Time: Evidence from Auctioned Cab Rides." *Revise & Resubmit, Econometrica*.

Cervero, Robert, and Aaron Golub. 2007. "Informal transport: A global perspective." *Transport policy* 14 (6): 445–457.

City of Cape Town. 2014. *Operating Licence Strategy 2013-2018.* Technical report. City of Cape Town Transport and Urban Development Authority, October. https://tdacontenthubstore.blob.core.windows.net/resources/53226657-22e8-4795-b9f8-144f2b535636.pdf.

Coetzee, Justin, Christoff Krogscheepers, and John Spotten. 2018. "Mapping minibus-taxi operations at a metropolitan scale - methodologies for unprecedented data collection using a smartphone application and data management techniques."

Collard-Wexler, Allan, Gautam Gowrisankaran, and Robin S. Lee. 2019. "Nash-in-Nash" Bargaining: A Microfoundation for Applied Work." *Journal of Political Economy* 127 (1): 163–195.

Combes, Pierre-Philippe, and Laurent Gobillon. 2015. "The Empirics of Agglomeration Economies." In *Handbook of Regional and Urban Economics,* vol. 5A, 247–348. Elsevier.

Commission, Competition. 2020. *Market Inquiry into Land Based Public Passenger Transport.* Technical report. Competition Commission. https://www.compcom.co.za/wp-content/uploads/2020/02/PROVISIONAL-MAIN-REPORT-FOR-PUBLIC-COMMENT_-19February2020-NON-CONFIDENTIAL-VERSION.pdf.

De Vos, Jonas, Alireza Ermagun, and F. Atiyya Shaw. 2023. "Wait time, travel time and waiting during travel: existing research and future directions." *Transport Reviews* 43 (5): 805–810.

Donald, Eric, Masao Fukui, and Yuhei Miyauchi. 2025. "Optimal Dynamic Spatial Policy." *Revise & Resubmit, Journal of Political Economy*.

Fabre, Anais, and Stephane Straub. 2023. "The Impact of Public–Private Partnerships (PPPs) in Infrastructure, Health, and Education." *Journal of Economic Literature* 61 (2): 655–715.

Gechter, Michael, and Nick Tsivanidis. 2023. "Spatial Spillovers from High-Rise Developments: Evidence from the Mumbai Mills." *Revise & Resubmit accepted, Econometrica*.

Goldszmidt, Ariel, John A. List, Robert D. Metcalfe, Ian Muir, V. Kerry Smith, and Jenny Wang. 2020. "The Value of Time in the United States: Estimates from Nationwide Natural Field Experiments."

Harari, Mariaflavia, and Maisy Wong. 2024. "Slum Upgrading and Long-run Urban Development: Evidence from Indonesia." *Accepted, Review of Economic Studies*.

Hillier, Frederick Stanton, and Gerald J. Lieberman. 2024. "Introduction to operations research." Chap. 17, 777–846. New York: McGraw-Hill Education.

Hosios, Arthur J. 1990. "On The Efficiency of Matching and Related Models of Search and Unemployment." *The Review of Economic Studies* 57 (2): 279–298.

Hsiao, Allan. 2023. "Sea Level Rise and Urban Adaptation in Jakarta."

———. 2024. "Educational Investment in Spatial Equilibrium: Evidence from Indonesia."

International Comparison Program, World Bank. 2020. *World Development Indicators database.* Technical report. World Bank and Eurostat-OECD PPP Programme.

Jobanputra, Rahul. 2018. *Comprehensive Integrated Transport Plan 2018 – 2023.* Technical report. City of Cape Town Transport and Urban Development Authority, January. https://tdacontenthubstore.blob.core.windows.net/resources/fd3ddc0d-b459-4d26-bb01-7f689d7a36eb.pdf.

Johnston, Robert J., Kevin J. Boyle, Wiktor (Vic) Adamowicz, Jeff Bennett, Roy Brouwer, Trudy Ann Cameron, W. Michael Hanemann, et al. 2017. "Contemporary Guidance for Stated Preference Studies." *Journal of the Association of Environmental and Resource Economists* 4 (2): 319–405.

Kerr, Andrew. 2018. *Background note: Minibus Taxis, Public Transport, and the Poor.* Technical report. World Bank. https://openknowledge.worldbank.org/handle/10986/30018.

Kerzhner, Tamara. 2022. "Is informal transport flexible?" *The Journal of Transport and Land Use* 15 (1): 671–689.

———. 2023. "How are informal transport networks formed? Bridging planning and political economy of labour." *Cities* 137:104348.

Khanna, Gaurav, Carlos Medina, Anant Nyshadham, Daniel Ramos-Menchelli, Jorge Tamayo, and Audrey Tiew. 2024. "Spatial Mobility, Economic Opportunity, and Crime." *Revise & Resubmit, American Economic Review*.

Kreindler, Gabriel E. 2022. "Peak-Hour Road Congestion Pricing: Experimental Evidence and Equilibrium Implications." *Conditionally accepted, Econometrica.*

Kreindler, Gabriel E., Arya Gaduh, Tilman Graff, Rema Hanna, and Benjamin A. Olken. 2023. "Optimal Public Transportation Networks: Evidence from the World's Largest Bus Rapid Transit System in Jakarta."

Mangham, Lindsay J., Kara Hanson, and Barbara McPake. 2009. "How to do (or not to do)...Designing a discrete choice experiment for application in a low-income country." *Health Policy and Planning* 24:151–158.

Mbonu, Oluchi, and F. Christopher Eaglin. 2024. "Market Segmentation and Coordination Costs: Evidence from Johannesburg's Minibus Networks."

Miyauchi, Yuhei, Kentaro Nakajima, and Stephen J. Redding. 2022. "The Economics of Spatial Mobility: Theory and Evidence Using Smartphone Data."

Mohring, Herbert. 1972. "Optimization and Scale Economies in Urban Bus Transportation." *American Economic Review* 62 (4): 591–604.

Moovit. 2024. *Moovit Global Public Transport Report 2024.* Technical report. Moovit. https://moovitapp.com/report.

Nikolaidou, Anastasia, Aristomenis Kopsacheilis, Georgios Georgiadis, Theodoros Noutsias, Ioannis Politis, and Ioannis Fyrogenis. 2023. "Factors affecting public transport performance due to the COVID-19 outbreak: A worldwide analysis." *Cities* 134:104206.

Norman, Peter. 2004. "Efficient Mechanisms for Public Goods with Use Exclusions." *The Review of Economic Studies* 71 (4): 1163–1188.

OECD. 2016. *Time spent travelling to and from work.* Technical report LMF2.6. OECD Social Policy Division - Directorate of Employment, Labour and Social Affairs, December. https://www.oecd.org/els/family/LMF2_6_Time_spent_travelling_to_and_from_work.pdf.

OECD/SWAC. 2023. *Mapping Territorial Transformations in Africa (MAPTA) Urban Form Dataset.* Technical report. OECD/SWAC. https://mapping-africa-transformations.org/climate.

Olea, José Luis Montiel, and Carolin Pflueger. 2013. "A Robust Test for Weak Instruments." *Journal of Business & Economic Statistics* 31 (3): 358–369.

Parry, Ian W. H., and Kenneth A. Small. 2009. "Should Urban Transit Subsidies Be Reduced?" *American Economic Review* 99 (3): 700–724.

Rose, John M., and Michiel C.J. Bliemer. 2009. "Constructing efficient stated choice experimental designs." *Transport Reviews* 29 (5): 587–617.

Schalekamp, H., and J. M. Klopp. 2018. "Beyond BRT: Innovation in Minibus-Taxi Reform in South African Cities."

Schalekamp, Herrie, and Nino McLachlan. 2016. "Minibus-taxi operator reforms, engagement and attitudes in Cape Town." In *Paratransit in African Cities: Operations, Regulation and Reform,* edited by Roger Behrens, Dorothy McCormick, and David Mfinanga, 174–198. Routledge.

Small, Kenneth A., and Erik T. Verhoef. 2007. *The Economics of Urban Transportation.* New York: Routledge.

Spence, A. Michael. 1975. "Monopoly, Quality, and Regulation." *The Bell Journal of Economics* 6 (2): 417–429.

Tsivanidis, Nick. 2023. "Evaluating the Impact of Urban Transit Infrastructure: Evidence from Bogotá's TransMilenio." *Conditionally accepted, American Economic Review*.

Tun, Thet Hein, and Darío Hidalgo. 2022. *Learning Guide: Toward Efficient Informal Urban Transit.* Technical report. WRI Ross Center for Sustainable Cities and Transformative Urban Mobility Initiative (TUMI). https://thecityfixlearn.org/en/learning-guide/toward-efficient-informal-urban-transit.

Vitali, Anna. 2024. "Consumer Search and Firm Location: Theory and Evidence from the Garment Sector in Uganda."

Whittington, Dale. 2010. "What Have We Learned from 20 Years of Stated Preference Research in Less-Developed Countries?" *Annual Review of Economics* 2:209–236.

Woolf, S.E., and J.W. Joubert. 2013. "A people-centred view on paratransit in South Africa." *Cities* 35:284–293.

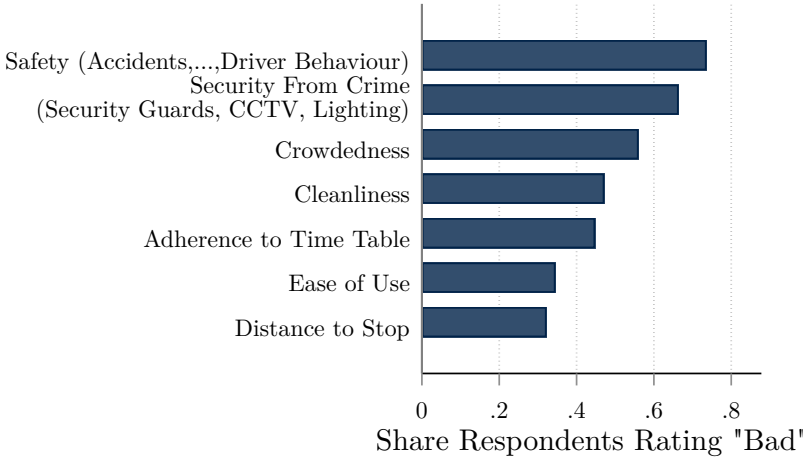
Yürükoğlu, Ali. 2022. "Empirical Models of Bargaining with Externalities in IO and Trade." In *Bargaining,* edited by Emin Karagözoğlu and Kyle B. Hyndman, 227–247. Palgrave Macmillan, Cham.

Zarate, Roman David. 2024. "Spatial Misallocation, Informality, and Transit Improvements: Evidence from Mexico City." *Revise & Resubmit, American Economic Review*.

Online Appendix

A. Additional Figures and Tables

Figure A.1. Ratings of Minibus Attributes

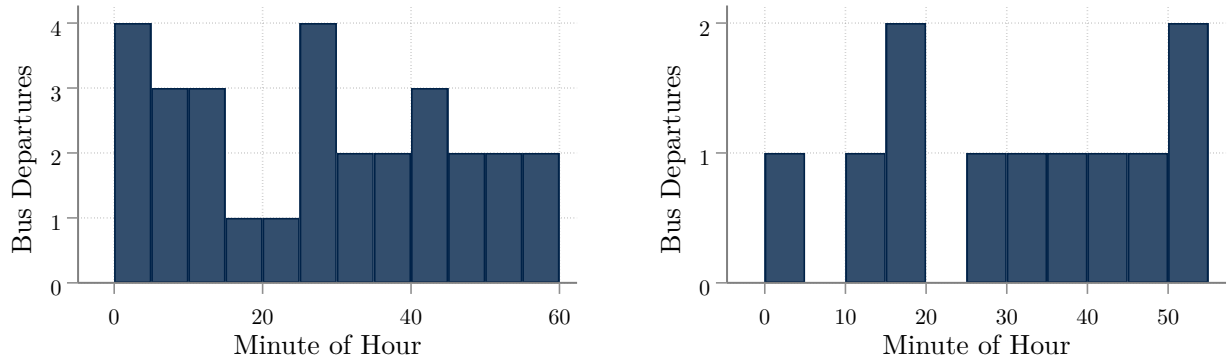


Notes: Figure displays a bar graph of the shares of respondents (N = 1685) in the 2013 Cape Town Household Travel Survey rating each minibus attribute as "bad," as opposed to "acceptable" or "good."

Figure A.2. Bus Departure Times

(a) Medium-Demand Route: Mitchells Plain-CBD

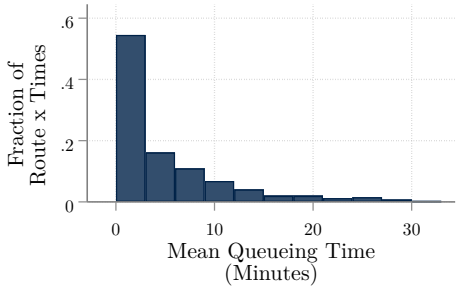
(b) Low-Demand Route: Wesbank-Eerste River



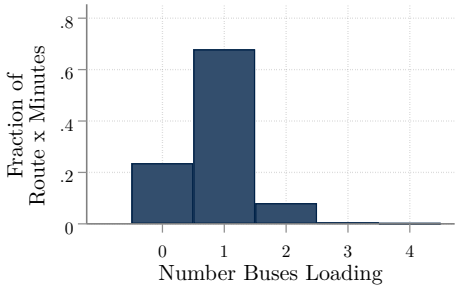
Notes: Panel (A) displays the distribution of minutes of the hour at which buses depart on a route, Mitchells Plain-CBD, with approximately median passenger demand, based on my station count data. Panel (B) displays the same distribution for a route, Wesbank-Eerste River, with passenger numbers at the 25th percentile across routes.

FIGURE A.3. MINIBUS LOADING PROCESS: DISTRIBUTIONS OF. . .

(A) QUEUEING TIME



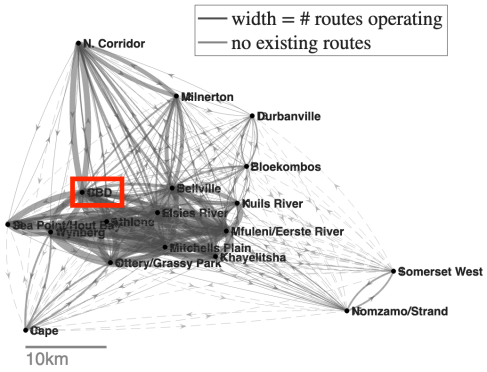
(B) BUSES LOADING SIMULTANEOUSLY



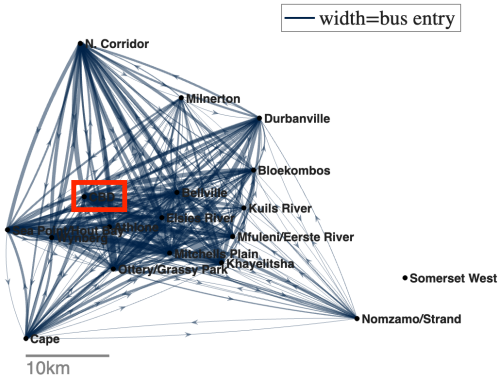
Notes: Panel (A) displays the distribution of passenger queueing times at the origin station for a specific route, over minibus routes and five-minute periods in my station count data from Cape Town. Panel (B) displays the distribution of the number of minibuses loading simultaneously at the origin station on a specific minibus route, over minibus routes and minutes.

FIGURE A.4. MINIBUS NETWORK IN DATA VERSUS MODEL

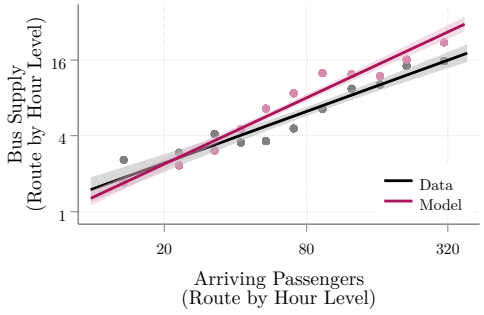
(A) NUMBER ROUTES, DATA



(B) MODEL-PREDICTED BUS SUPPLY, b_{ij}



(C) BUS SUPPLY VERSUS PASSENGERS IN DATA VERSUS MODEL



Notes: Map in Panel (A) displays the number of minibus routes linking each origin-destination pair of transport analysis zones according to a GIS shapefile created through a collaboration between GoMetro and the City of Cape Town. Note that, since these neighborhood units include multiple minibus stations in the real world, many pairs are linked by multiple "routes" in the data, in contrast to my model. The map in Panel (B) displays origin-to-destination lines whose thickness corresponds to the model-predicted minibus supply, b_{ij}, on that origin-destination route. Both maps highlight the CBD in red. Panel (C), in turn, displays binned scatterplots and best-fit lines of the log-scale relationship between bus supply, b_{ij}, and newly-arriving passengers per hour, proportional to λ_{ij}, across routes and hours in the station count data and across routes as predicted by the model.

B. Data

B.1 Minibus Station Counts

In this section, I provide more detail on the minibus station counts used to characterize the matching process. I designed the counts in cooperation with the mobility advisory firm GoAscendal, which also organized the logistics of data collection.

B.1.1 Sample

Resources allowed for the enumeration of 6 minibus routes per minibus station at 8 stations. For supervisory purposes, routes had to be enumerated in groups of 6, all operating from the same station.

Sampling Frame Complicating the sampling procedure is the fact that no fully comprehensive, accurate list of stations and routes exists. Thus, as a sampling frame, I employ a roster of routes by origin station derived from a 2018 collaboration between GoMetro and the City of Cape Town's Transport and Urban Development Authority.[52] This listing is, according to stakeholders, as comprehensive and accurate as any available and includes the number of minibus trips mapped by route in this previous data collection effort. Stakeholder discussions revealed that the number of trips mapped is an indicator of the number of buses on a route.

Population Since the team of 12 enumerators, who could cover 6 routes, had to be employed at the same station on a given day, my survey population consists of minibus routes originating from stations in Cape Town with at least 6 routes. The aforementioned sampling frame lists 107 stations. Of these stations, 31 have at least 6 routes, and these stations, as well as the 328 routes which originate from them, enter the first and second stages of my sampling frame, respectively, as described in the next section.

Clusters and Stratification I employed a two-stage stratified cluster sample, sampling 8 stations and then 6 routes originating at each of the 8 stations. I stratify within each stage by a proxy for station and route-level bus supply, namely the aforementioned number of trips mapped in the previous data collection effort, to over-sample stations and routes with higher levels of bus supply and thus reduce the number of zero bus and passenger observations in the resulting data.

In the first stage, I took a stratified random sample of stations. First, I took a 100% sample of the 5 highest-bus-supply feasible stations that operate in the morning peak, as measured by my proxy of bus supply, i.e. the total trips mapped originating at that station in the previous 2018 study.[53] Second, I sampled 3 stations, or a 16% sample, from the remaining 19 non-duplicate lower-bus-supply stations with no permission issues,

[52] In order to update the city's record of on-the-ground minibus route paths and operations, GoMetro sent enumerators armed with a smartphone app to ride on close to 30,000 minibus trips on the approximately 800 established minibus routes. The results showed the official, city-designated routes to be outdated. For example, 250 of the official city routes no longer operated (Coetzee et al. (2018)).

[53] From the full listing of 31 stations, some which would have otherwise counted among the 5 busiest had to be skipped due to minibus associations denying permission (Nyanga Central, Gugulethu Eyona) or not containing 6 routes that load off-road (Claremont Station). The 5 busiest feasible stations are numbers 1-2, 4, 6, and 8.

redrawing stations until obtaining 2 feasible ones. The final "busy" sample includes Du Noon, Bellville, Mfuleni, Khayelitsha Site C, and Mitchells Plain Station Eastern Side (North); the "less busy" sample includes Elsies River, Wesbank, and Nomzamo.[54]

In the second stage, I took a stratified random sample of routes within each cluster, or station. Specifically, I sampled 4 routes (or the maximum possible, up to 4) per station from among those in the top ten percent of bus supply, as measured by trips mapped in the previous study, across all routes serving the 8 sampled stations. Then, I draw the remaining two or more routes, for a total of six, from those below the top ten percent of trips mapped. Thus, while feasibility constraints do not permit a constant sampling rate across stations, I obtain a random sample with variation in the traffic levels across routes. My final two-stage cluster sample of routes thus contains 48 routes clustered across six stations.

Final Sample vs. Population In the field, 14 sampled routes unexpectedly did not operate at all in the morning peak. Field supervisors then randomly selected replacement routes at the same station on the spot, to the extent that additional routes were operating at the station. One station, Elsies River, was discovered to have only 2 total routes in operation upon commencement of the day's data collection, and logistical considerations meant that no replacement routes at another station could be chosen. As a result, my final sample comprises 44 rather than 48 routes.

I now endeavor to compare the characteristics of routes in my sample with those of the entire population of routes in Cape Town. Note that I can do so only for 41 of my 44 sampled routes. Three routes in my sample, which were randomly chosen to replace infeasible or nonexistent originally-sampled routes, do not appear in the original sampling frame, so I lack data for these routes on the covariates which would allow comparison to the population of routes. For ease of interpretation, I now merge both directions of routes as well as any duplicate routes with the same origin and destination to obtain a population of approximately 429 unique routes in Cape Town.[55]

Thus, in Online Appendix Table B.1, I compare the 41 in-sample routes that are also in the original sampling frame to the universe of 429 unique routes in Cape Town and find that both the "busy" routes with bus supply above the threshold used for the stratification and the "other," less-busy routes are broadly representative of the respective populations in terms of bus supply, as measured by the number of trips recorded in the aforementioned 2018 roster, length, and fares paid.

Comparability of Routes Since I use the station count data to demonstrate relationships between demand and other outcomes across routes, I now explore the extent to which commuters on routes with different levels of demand are comparable in terms of observables. In Online Appendix Table B.2, I compare commuter

[54] Wynberg Station (Western Side) had to be excluded from the less busy sampling frame due to lack of permission, Cape Town CBD station due to lack of AM peak operations, and Mitchell's Plain Station (North) and Promenade as well as Mitchells Plain Station Eastern Side (South) due to being adjacent to an already sampled station. After drawing the sample, further stations had to be excluded and resampled as follows. One station sampled, Khayelitsha (Vuyani), turned out to be part of an already sampled station (Khayelitsha (Nolungile Site C)); several others (Athlone, Vasco Station) do not operate as minibus stations with queues and loading off-road, and another set (Zevenwacht Mall, Mitchells Plain (Promenade), Tableview (Bayside)) do not operate in the AM peak.

[55] Note that, unlike in my sampling frame, I do not exclude from this comparison population the stations with fewer than 6 routes.

TABLE B.1. STATION COUNT SAMPLE CHARACTERISTICS

	Busy Routes		*Other Routes*	
Variable	Sample	All Routes	Sample	All Routes
Mean # Trips Recorded	195.5	196.5	52.4	55.0
Mean Trip Length (km)	15.6	13.0	17.1	12.1
Mean Fare	11.5	10.9	12.8	11.1
N Routes	17	79	24	350

Notes: This table displays characteristics for 41 of the routes included in my 44-route minibus station count sample, as compared to the universe of approximately 429 unique routes in a roster developed through a 2018 collaboration between GoMetro and the City of Cape Town. Note that I merge both directions of routes, when these are recorded separately, as well as any duplicate routes with the same origin and destination. Three routes in my 44-route sample, which replaced infeasible or nonexistent sampled routes, were not present in the roster and so are excluded from the table. The number of trips refers to those recorded in the roster, which is a proxy for bus supply. "Busy" routes are those where the trips recorded lie above the 90th percentile across routes serving my sampled stations.

characteristics across origin-destination pairs of transport analysis zones with above- versus below-median total daily passenger arrivals on routes in my station count sample. I find that commuters on higher-demand routes are broadly similar to those on lower-demand routes in terms of observable characteristics, including gender, education, age, home and work distance from the CBD, and time leaving home to go to work. However, commuters on higher-demand routes have somewhat lower median incomes. Both above- and below-median demand routes have similar fares and length.

B.1.2 Data Collected

Station counts occurred on weekday mornings during one morning peak period (6-10am) per station, on weekdays from June 20-28 and 30, 2022. Two enumerators recorded data on each of 6 sampled routes over the course of the four-hour period. One enumerator stationed at the beginning of the loading lane and passenger queue corresponding to a route recorded, on forms as in Online Appendix Figure B.5a, the time a minibus vehicle arrived to the station and, every 5 minutes, the queue length, i.e. the number of passengers waiting in the queue for that route.[56] The second enumerator per route monitored bus loading and departures, recording the time a vehicle pulled in to the front of the loading lane where buses typically load passengers, the time of departure, as well as the number of passengers on board at departure, as in Online Appendix Figure B.5b.

B.1.3 Calculations

First, I calculate quantities related to the minibus loading process for the facts in Section III. I begin by recording the number of buses at the front of the loading bay, where passengers typically board, for each route and minute. I then discretize time into 5-minute periods, each beginning at some clock time t. I measure loading time L_{sijt} as the number of minutes between the time the minibus for trip s on route ij departing during time interval t arrived to the front of the loading bay and the time it departs, as described in Online

[56] Enumerators were instructed to count only the passengers waiting in the queue for that route exactly at each five-minute mark. In other words, if, at a given time, all passengers directly walk onto loading buses without having to queue, queue length equals zero, consistent with the queueing model which I later employ this data to estimate.

Table B.2. Minibus Commuter and Route Characteristics By Route Demand

	Total Daily Passenger Demand...	
Variable	Below-Median	Above-Median
Commuter Characteristics: Share...		
Auto Owners	0.272	0.358
Female	0.535	0.567
College-Educated	0.0521	0.0381
Living in Shack	0.102	0.177
Commuter Characteristics: Median...		
Monthly Personal Income [using bin midpoints]	$191	$143
Age	33	34.50
Household Size	4	4
Distance, Home Location to CBD	19.39	18.63
Distance, Work Location to CBD	15.64	17.53
Hour Leaving for Work	6	6
Route Characteristics: Median...		
Fare ($)	1.186	1.121
Origin-Destination Distance (km)	12.79	12.68

Notes: This table displays means and medians of demographic characteristics of minibus commuters across routes in my station count data, separately for routes with above- versus below-median total daily minibus passenger arrivals, as well as median characteristics of these routes, including fares and straight-line distance from origin to destination. Demographic characteristics are measured in the 2013 Cape Town Household Travel Survey. I aggregate both datasets to origin-destination pairs of transport analysis zones (TAZs).

Appendix Section B.1.2. I assume that the passengers I observe departing from the origin station on trip s, deppax_{sijt}, board the bus at a uniform rate. Then, I proportionately apportion these passengers who depart on a trip to the five-minute blocks during which the bus was loading to calculate the total number of passengers boarding buses on route ij from t to $t+5$ as

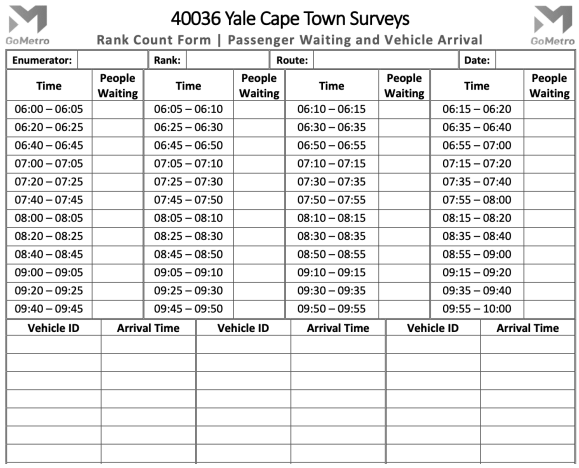
$$\text{loading passengers}_{ijt} \equiv \sum_{s} \frac{L_{sijt} \cap [t, t+5)}{L_{sijt}} \text{deppax}_{sijt}.$$

Here, $L_{sijt} \cap [t, t+5)$ indicates the number of minutes of trip s's loading time that overlap temporally with clock times t to $t+5$. I observe the queue length in passengers, n_{ijt}, at the beginning of every 5-minute block and then calculate the number of newly arriving passengers λ_{ijt} per minute as

$$\lambda_{ijt} \equiv \frac{n_{ij,t+1} + \text{loading passengers}_{ijt} - n_{ijt}}{5}$$

FIGURE B.5. STATION COUNT DATA COLLECTION FORMS

(A) PASSENGER QUEUES



40036 Yale Cape Town Surveys

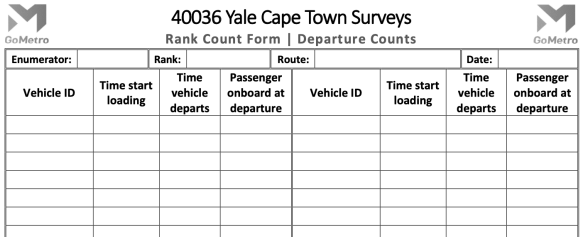
Rank Count Form | Passenger Waiting and Vehicle Arrival

Enumerator:		Rank:		Route:		Date:	

Time	People Waiting	Time	People Waiting	Time	People Waiting	Time	People Waiting
06:00 – 06:05		06:05 – 06:10		06:10 – 06:15		06:15 – 06:20	
06:20 – 06:25		06:25 – 06:30		06:30 – 06:35		06:35 – 06:40	
06:40 – 06:45		06:45 – 06:50		06:50 – 06:55		06:55 – 07:00	
07:00 – 07:05		07:05 – 07:10		07:10 – 07:15		07:15 – 07:20	
07:20 – 07:25		07:25 – 07:30		07:30 – 07:35		07:35 – 07:40	
07:40 – 07:45		07:45 – 07:50		07:50 – 07:55		07:55 – 08:00	
08:00 – 08:05		08:05 – 08:10		08:10 – 08:15		08:15 – 08:20	
08:20 – 08:25		08:25 – 08:30		08:30 – 08:35		08:35 – 08:40	
08:40 – 08:45		08:45 – 08:50		08:50 – 08:55		08:55 – 09:00	
09:00 – 09:05		09:05 – 09:10		09:10 – 09:15		09:15 – 09:20	
09:20 – 09:25		09:25 – 09:30		09:30 – 09:35		09:35 – 09:40	
09:40 – 09:45		09:45 – 09:50		09:50 – 09:55		09:55 – 10:00	

Vehicle ID	Arrival Time	Vehicle ID	Arrival Time	Vehicle ID	Arrival Time

(B) BUS LOADING AND DEPARTURE



40036 Yale Cape Town Surveys

Rank Count Form | Departure Counts

Enumerator:		Rank:		Route:		Date:	

Vehicle ID	Time start loading	Time vehicle departs	Passenger onboard at departure	Vehicle ID	Time start loading	Time vehicle departs	Passenger onboard at departure

Notes: This figure displays the data collection forms used by enumerators to record station count data by hand for later digitization. Form (A) was used by the first of two enumerators to record the length of the passenger queue on an assigned route every 5 minutes from 6-10am, as well as each minibus that arrived on the station premises and its time of arrival. Form (B) was used by the second enumerator to record, for every minibus that loaded passengers between 6-10am on a given route, the time it pulled in to the front of the loading bay ("Time start loading"), the time of departure from the station, and the number of passengers on-board at departure.

under the assumption that no passengers abandon the queue after joining. I display these two variables, n_{ijt} and $\lambda_{ij,t-1}$, in Figure 7a.

The mean queueing time of these passengers to board buses, consistent with an approximately stationary queue within some temporal neighborhood of 5-minute time interval t, is $Q_{ijt} \equiv 5 \cdot \frac{n_{ijt}}{\text{loading passengers}_{ijt}}$ and is measured in minutes. For Figures 6a-6b and 7b, I then calculate route-by-hour h-by-date d averages of L_{sijt}, Q_{ijt}, and λ_{ijt} – where, recall, t denotes 5-minute intervals – which I denote by L_{ijhd}, Q_{ijhd}, and λ_{ijhd}. Average total wait time then equals $W_{ijhd} = Q_{ijhd} + L_{ijhd}$. For Figure 7b, I also use the vehicle IDs of buses observed arriving to the station on a given route during a given hour h on date d to calculate the number of unique buses b_{ijhd} supplied to the route.

Second, I detail the quantities needed to estimate queueing efficiency μ, as in Equation (13). I already have bus-departure-level loading times L_{sijt}, which, additionally indexed by the date d within which a 5-minute block t falls, are the L_{sijtd} on the left-hand side of (13). I then calculate the probability of the existence of a (nonzero) queue of passengers, conditional on a bus being present, or "queue prevalence" for short. To calculate queue prevalence p_{ijtd} for each route ij as well as five-minute time period t and date d, I calculate, within a one-hour rolling window centered on t, the share of five-minute periods at the start of which I see one or more passengers in the queue *and* a bus present in the loading area. To obtain the conditional probability p_{ijtd}, I then divide by the total share of time periods t within this hour window at the start of which a bus is present.

Third, I calculate variables for estimation of the minibus arrival efficiency ρ and internal calibration. Turning

first to the estimation of arrival efficiency in (14), I have already discussed the calculation of route-by-hour-by date bus supply b_{ijhd} and trip-s-level loading times L_{sijhd}, indexed by route as well as the hour h and date d during which the bus departs. I then calculate, for minibus trip (i.e. departing bus) s, the minutes elapsed between the departure of the previous bus and the instant when bus s arrives to the front of the loading bay, which yields gap_{sijhd}. Using my separate onboard tracking data, I calculate the average origin-destination travel time on route ij, which yields T_{ij}. For internal calibration, I calculate the median minibus fare by first taking a passenger-weighted average by route in the same onboard tracking data and then taking the median across routes. The median queue length in the internal calibration equals the median of every-five-minute queue length observations n_{ijt}.

B.2 Minibus Stated Preference Survey

Next, I detail my stated preference survey, also implemented by the mobility advisory firm GoAscendal.

B.2.1 Questionnaire Design

I designed the questionnaire to maximize statistical power while retaining respondent attention. Since the 2013 Cape Town Household Travel Survey already contains discrete choice experiments containing different modes of transport, I focus exclusively on minibus commutes with different non-pecuniary attributes and costs.

Choice of Attributes and Levels In a discrete choice experiment, attributes should be chosen that are important and relevant to the decision at hand (Mangham et al. (2009) and Johnston et al. (2017)). Conveniently, the aforementioned Cape Town Household Travel Survey asks respondents to rate the importance of a variety of factors in their mode choice decisions; the three factors most frequently rated "most important" in mode choice were comfort, safety, and security. In separate questions asking respondents to rate various aspects of existing minibus service on a four-point scale, "Safety (accidents, maintenance, driver behavior)," "Security from crime," and "Availability of a seat/crowdedness" are also those most frequently rated "bad." [57]

I thus choose three nonpecuniary attributes, or "quality improvements," corresponding to mode users' three main concerns: the presence or absence of security guards, driver adherence to speed limits, and whether the minibus loads more passengers than seats. Additionally, I stipulate a travel time and cost (fare) for each minibus alternative. In line with guidance in the literature (Johnston et al. (2017) and Mangham et al. (2009)), I choose attribute levels for the quantitative attributes that are plausible and within the range of typically experienced values in Cape Town yet allow for sufficient variation.[58]

D-Efficiency Algorithm I use the Stata package `dcreate` to choose a questionnaire design, namely the combinations of attribute levels in each alternative of each choice set presented to respondents. This *d-efficiency* algorithm minimizes the determinant of the variance-covariance matrix of the estimated parameters

[57] Other aspects included timetable adherence, cleanliness, distance to stop, and ease of use.

[58] Fares can take values of R6, R10, R14, and R18, while travel time is either 20, 30, 40, or 50 min., corresponding to the lengths of typical minibus rides in the morning peak.

of a discrete choice model for given coefficient priors (Ben-Akiva et al. (2019) and Rose and Bliemer (2009)). In addition, I specified, in the introductory script, a wait time of 10 minutes, which is constant across all choice sets and alternatives. As the discrete choice model whose statistical power is maximized, I use a version of my model where passengers pay a flow utility cost only while traveling on a minibus, rather than a one-time utility cost.

Coefficient Priors To implement the d-efficiency algorithm and thereby choose attribute levels for the final survey, I now require priors for the demand model parameters. I obtain priors $\nu = 12.7$, $r = 0.002$, and $\kappa_M = 0.88$ from estimating a mode choice model on the Cape Town household travel survey stated preference module.[59] Then, since no larger-sample stated preference surveys cover analogous attributes, I use the results of my 1-day stated preference pilot survey–which surveyed $N = 20$ commuters using a very similar format–to estimate priors for ξ_z, $z \in \{\text{security}, \text{no speeding}, \text{no overloading}\}$, obtaining $\xi_{\text{security}} = -0.18$, $\xi_{\text{no speeding}} = -0.16$, and $\xi_{\text{no overloading}} = -0.21$.[60] I set $\xi_z = -0.1$ for each z since the pilot estimates are noisy and use the median household income per working day from the Cape Town Household Travel Survey to set $\omega_i = 427$.

Questionnaire Dimensions I employ these priors in the d-efficiency algorithm to generate 2 "blocks," or versions, of 5 choice sets with 2 alternatives each.[61] The 6th choice set, common to both blocks, had a strictly dominant option and thus could provide a measure of comprehension. I do not include an outside option (Ben-Akiva et al. (2019)), as my survey is intended to test relative, rather than absolute, demand for different minibus options; my quantitative model will yield the overall demand for minibus commuting. Furthermore, all attributes have pictograms to aid comprehension in a lower-education context (Mangham et al. (2009)). In addition, I collected demographic information: education, gender, age, personal income, and car ownership. I also collected transport-related information such as current trip purpose, usual commute modes, and frequency of minibus use.

B.2.2 Pilot Survey Lessons

The enumerator team and I conducted a pilot survey at the Cape Town CBD minibus station on 15 June 2022 from approximately 11am to 1pm, where we contacted 36 respondents, 25 of whom qualified for and completed the pilot questionnaire, which had one version (block) with 9 choice sets of 3 alternatives each and a second block with 9 choice sets of 2 alternatives each. Anecdotally, the respondents I interviewed seemed to be taking the scenarios seriously and understanding the aim of the exercise, musing out loud, for example, "I can't take this bus because it will make me late to work!" However, after the pilot survey, I reduced the

[59] I restrict the sample to choice sets that do not contain car as a mode and to respondents aged 25-65 who work outside the home.

[60] In estimating the multinomial logit on pilot data, I restrict the coefficients on travel time, cost, and the travel-time income interaction to be consistent with the aforementioned two priors and use the midpoints of household income bins from a separate income question.

[61] I create two blocks, which are randomized across respondents, because doing so increased power in Monte Carlo simulations without increasing respondent burden. As for the numbers of choice sets and alternatives, I reduced these from 8 to 5 and 3 to 2, respectively, after the pilot revealed respondent frustration and inattention towards the end of the survey – and a version with 2 rather than 3 alternatives proved less problematic in this regard.

number of choice sets and alternatives per choice set to maintain respondent attention.

B.2.3 Sample

Stated preference surveys were conducted at one mall and intermodal transport interchange and two minibus stations, for 5 weekday hours per location (11am-4pm) on 21, 27, and 30 June 2022. Security considerations did not permit a random sampling of minibus stations or other locations, as many were not deemed safe to approach strangers for this kind of survey. At the Middestad Mall/Bellville transport interchange, enumerators were instructed to conduct surveys inside the mall, at the Golden Arrow public bus station, and on surrounding streets, but explicitly *not* within the minibus station. On the other hand, at the Khayelitsha Site C and Somerset West Shoprite minibus stations, interviews were conducted only within the minibus station. The aim here was to obtain, at the mall and transport interchange, a representative sample of different mode users as well as, at the minibus stations, a sample of respondents intimately familiar with minibuses, for whom the hypothetical alternatives would be similar to their existing commutes. Only (full-, part-time, or self-) employed respondents were interviewed, so that the scenarios correspond to my quantitative model.[62]

B.2.4 Administration and Script

Survey enumerators randomly approached respondents and asked for their consent to participate in the survey, according to a script in Online Appendix Figure B.6a. They were offered a chocolate as an incentive. Enumerators then proceeded to read them questions shown in the Survey CTO Android app (see Online Appendix Figures B.6b-B.6d), which automatically progresses through the questionnaire, showing follow-up questions or terminating the survey where appropriate. The stated preference scenarios themselves were shown on laminated paper, and enumerators also entered responses directly into the app.

B.2.5 Field Experience

All 526 employed respondents who began the survey also completed all questions. Note that there were no corner solutions: every alternative of every question was chosen by a nonzero number of respondents.

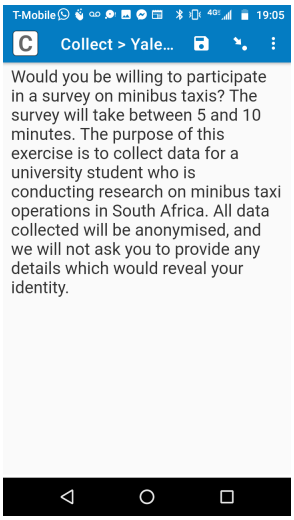
B.2.6 Sample Characteristics

In later estimation, I stack my own stated preference survey with the stated preference module of the 2013 Cape Town Household Travel Survey. Online Appendix Table B.3 compares the demographic characteristics of each sample to the aggregate city commuter population, as measured by the same 2013 survey. Along basic demographic dimensions, including gender, education, income, and age, both stated preference samples are representative of the aggregate population. However, respondents in my new sample are less likely to own cars, and, not surprisingly, given that many were recruited at minibus stations, more likely to report that they typically commute by minibus.

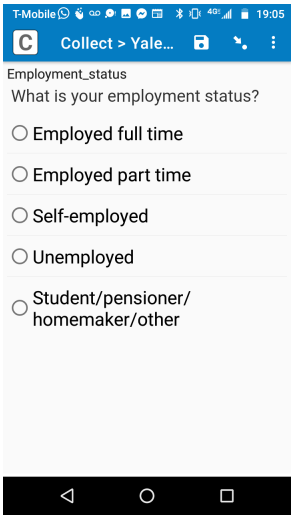
[62] Enumerators approached 586 people. Of these, 333 were full-time employed, 97 part-time employed, and 96 self-employed, for a total of 526 respondents who qualified for the survey. Of the remaining people not qualifying, 14 were students/pensioners/homemakers/other and 42 were unemployed.

FIGURE B.6. STATED PREFERENCE SURVEY APP SCREENSHOTS

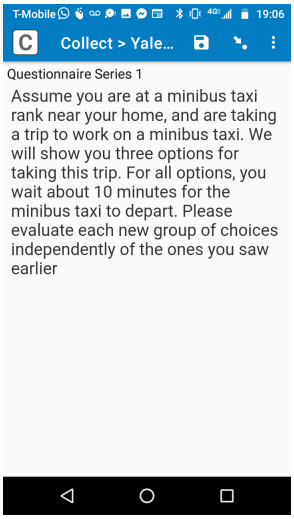
(A) CONSENT SCRIPT



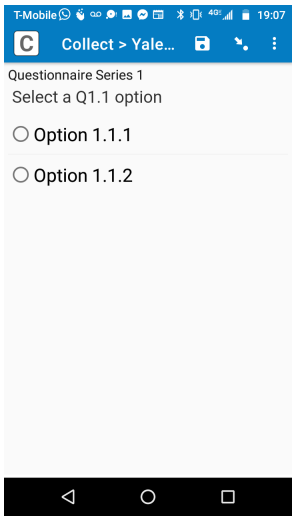
(B) DEMOGRAPHIC QUESTIONS



(C) PRE-SCENARIO SCRIPT



(D) RESPONSE ENTRY



Notes: These images show screenshots from the Survey CTO app used by enumerators to conduct and record stated preference responses, specifically (A) the consent script; (B) an example of a demographic question; (C) the script that introduces the stated preference choice sets; and (D) the screen used to enter stated preference responses.

In order to better understand whether this oversampling of minibus users might affect my results, Online Appendix Table B.4 displays demographic characteristics of commuters in the aggregate city population by their usual commute mode. Minibus users look similar to public bus or train commuters along most dimensions, though they work a bit farther from the central business district, on average. Relative to car commuters, minibus users are, not surprisingly, poorer and less-educated, both dimensions which will be captured by the model. Nonetheless, along other dimensions, such as the locations of their homes and jobs relative to the CBD, car commuters look more similar to minibus users. The fact that minibus users appear similar to other commuters along most dimensions excluded from my model provides initial reassurance that my sample could nonetheless produce estimates representative of the general population. Nonetheless, I later pursue multiple strategies to quantify any bias resulting from this oversampling of minibus users.

Furthermore, my choice of survey locations might have generated selected samples. To probe this possibility, in Table B.5, I compare demographic characteristics of respondents in my stated preference survey across interview location types: minibus stations, namely Khayelitsha Site C and Somerset West Shoprite, and non-minibus-station locations, namely the Bellville/Middestad Mall Transport Interchange. The minibus station sample is somewhat more-educated and has a larger female share, but the two groups of respondents otherwise look similar on observables, lending some confidence that selection on unobservables might not heavily bias my results.

TABLE B.3. STATED PREFERENCE SAMPLE CHARACTERISTICS

	Stated Pref. Samples		*Data*
Variable	Own	City-Run	Cape Town
Share Auto Owners	0.448	0.581	0.561
Share Female	0.458	0.494	0.458
Share College-Educated	0.295	0.228	0.190
Median Monthly Personal Income [using bin midpoints]	$287	$287	$287
Median Age	35	39	39
Commute Mode Shares of...			
Minibus	59.56	22.56	23.55
Public Bus/Train	19.61	27.69	22.81
Auto	12.11	40	39.40
Share Using Minibuses > 1x/Week	0.951	0.635	
N	413	407	

Notes: This table's first two columns display demographic characteristics of my newly-conducted stated preference survey sample as well as the 2013 Cape Town Household Travel Survey stated preference sample. The third lists the corresponding statistics in the aggregate Cape Town population, as inferred from a separate module of the latter survey. In each case, statistics reflect those samples used for estimation, namely respondents between the ages of 25 and 65 who work outside the home.

B.3 City of Cape Town Household Travel Survey (2013)

The City of Cape Town conducted the 2013 Cape Town Household Travel Survey (CTHHTS) on a representative sample of residents. In addition to demographics and car ownership, this survey records the addresses of residence and work, which I successfully geocode with the Google Geocoding API for $N = 17,395$ employed respondents, along with the details of respondents' commutes. For descriptive statistics and moments, I define the commute mode as follows: minibus includes any commuter who uses minibuses during his or her commute; public bus/train includes commuters who use train, bus, or MyCiti bus but not minibuses; auto includes car and motorcycle drivers and passengers who do not use minibuses or public bus/train; and the non-motorized and other category, all others.

A subset of respondents also completed a commute stated preference survey with a format similar to my own, except that respondents chose among different *modes* of transport. Respondents were asked to consider their last commute to work or education and then indicate their preferred mode option from each of 13 choice sets, of which Online Appendix Figure B.7 displays an example. Each choice set offered 3 mode options from among car, public MyCiti bus, (regular) public bus, public train, or minibus (taxi). Each option in a given choice set, in addition to representing a different mode, varied in monetary cost, travel time, waiting time, and number of transfers. However, this city-run survey did not include non-pecuniary quality improvements such as station security.

Table B.4. Commuter Characteristics by Transport Mode

Variable	Minibus	Public Bus/Train	Car
Share...			
Auto Owners	0.318	0.367	0.937
Female	0.571	0.535	0.434
College-Educated	0.0559	0.0804	0.368
Living in Shack	0.146	0.148	0.0188
Median...			
Monthly Personal Income [using bin midpoints]	$191	$287	$573
Age	35	36	39
Household Size	4	4	4
Distance, Home Location to CBD	18.22	18.95	14.74
Distance, Work Location to CBD	13.14	9.151	11.41

Notes: This table displays demographic characteristics of commuters by primary mode of transport as measured in the 2013 Cape Town Household Travel Survey.

Table B.5. Stated Preference Sample Characteristics by Interview Location

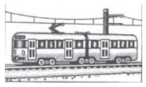
Variable	Minibus Station	Non-Minibus Station
Share Auto Owners	0.449	0.446
Share Female	0.489	0.396
Share College-Educated	0.328	0.230
Median Monthly Personal Income [using bin midpoints]	$287	$287
Median Age	36	34
Share Using Minibuses > 1x/Week	0.960	0.934
N	274	139

Notes: This table displays demographic characteristics of my newly-conducted stated preference survey sample by interview location, either the two minibus stations of Khayelitsha Site C and Somerset West Shoprite or the non-minibus station location, namely the Bellville/Middesstad Mall Intermodal Transport Interchange. In each case, statistics reflect those samples used for estimation, namely respondents between the ages of 25 and 65 who work outside the home.

B.4 Minibus On-Board Tracking Data

I make use of GPS-tracked minibus trips, also newly collected for this paper by the South African firm GoMetro. This data, logged by enumerators via a smartphone app, covers two trips from the beginning to the end of each route in my station count data and provides stop-level information within each trip. For each stop,

FIGURE B.7. EXAMPLE CITY OF CAPE TOWN STATED PREFERENCE SURVEY SCENARIO

Choice 1:	Train ☐	MyCiti Bus ☐	Minibus Taxi ☐
Cost	R 3	R 20	R 20
In- Vehicle Travel Time	70 min	25 min	70 min
Waiting time	15 min	20 min	10 min
Number of Transfers	2	2	2
No Preference ☐			

Notes: Figure displays example of choice set faced by respondents in the 2013 Cape Town Household Travel Survey, from which they indicated their preferred mode from among those listed based on the associated attributes.

I observe the number of passengers boarding and alighting, the arrival and departure time, and the fare paid by passengers boarding. In total, my sample includes $N = 582$ stops, made by 60 vehicles on 43 routes over 2 trips per route.

C. THEORY

C.1 Micro-Foundations of Commute Utility

I now lay out a micro-founded model of commuting which underlies the linear approximations to commute value U^g_{ijm} used in the main text. Suppose that commuters, upon birth, immediately enjoy their home-location-specific amenity θ^g_i, pay a one-time mode-specific utility cost κ^g_M and then wait for their chosen mode m, so their total (deterministic) utility satisfies $\overline{U}^g_{ijm} = \theta^g_i - \kappa^g_M + E\left[u^g_{ijm}\right]$ where the third term is the expected value of waiting, taken over the (possibly degenerate) distribution of wait times. Passengers board the vehicle after w_{ijm} minutes and immediately pay the fare τ_{ijm}. Given the rate of time preference r, the value of waiting, u^g_{ijm}, satisfies

$$u^g_{ijm} = \exp\left(-rw_{ijm}\right)\left(V^g_{ijm} - \tau_{ijm}\right), \tag{C.1}$$

and thus equals the value V^g_{ijm} of traveling by mode m from i to j, minus the fare, all discounted to account for wait time. The traveling value V^g_{ijm} equals the skill-specific wage ω^g_j received upon arrival, discounted to account for travel time t_{ijm}:

$$V^g_{ijm} = \exp\left(-rt_{ijm}\right)\omega^g_j, \tag{C.2}$$

Substituting (C.2) into (C.1) and taking a first-order approximation, around $r = 0$, to the value of waiting yields

$$u^g_{ijm} \approx \omega^g_j - \tau_{ijm} - r\omega^g_j\left(w_{ijm} + t_{ijm}\right) + r\tau_{ijm}w_{ijm} \tag{C.3}$$

Finally, taking expectations and substituting into total deterministic utility $\overline{U}^g_{ijm}$ yields

$$\overline{U}^g_{ijm} \approx \theta^g_i \underbrace{-\kappa^g_m - r\omega^g_j\left[E\left(w_{ijm}\right) + t_{ijm}\right] - \tau_{ijm}}_{\equiv U^g_{ijm}} + \omega^g_j. \tag{C.4}$$

where I have suppressed the term $r\tau_{ijm}E\left(w_{ijm}\right)$, which will be close to zero due to the multiplication of a small time preference rate and a small fare. I then define commute value U^g_{ijm} as the component corresponding to the costs of the actual commute. For the minibus mode, $m = M$, I set $E\left(w_{ijM}\right) = Q_{ij} + L_{ij}$, $t_{ijM} = T_{ij}$, and $\tau_{ijM} = \tau_{ij}$ to obtain Equation (7) in the main text. For public bus/train, $m = P$, $E\left(w_{ijP}\right) = H_{ij}$ and $t_{ijP} = T_{ijP}$, while, for the car mode, $m = A$, $E\left(w_{ijA}\right) = 0$, $t_{ijA} = T_{ij}$, and $\tau_{ijm} = \tau_A$.

C.2 Empirical Queueing Model

In this section, I introduce an empirical version of my queueing model with unobserved heterogeneity, which I omit from the main text for parsimony. I modify the queueing framework in Section IV in two ways. First, I assume that the bus arrival rate λ^B_{ij} additionally depends on an unobserved arrival speed ς_{ij} such that the expected gap between buses equals

$$\frac{1}{\lambda^B_{ij}} \equiv \frac{\log\left\{\exp\left[\rho\left(\frac{2T_{ij}}{b_{ij}} - L_{ij}\right)\right] + 1\right\} - \log\varsigma_{ij}}{\rho}. \tag{C.5}$$

Intuitively, ς_{ij} accounts for general congestion around minibus stations as well as unmeasured variations in station infrastructure. Second, suppose that, after a bus arrives at the origin station, the bus can load passengers only upon receipt of another (unobserved) "loading shock" at Poisson rate ι_{ij}, which captures idiosyncratic disruptions to the loading process, such as weather or special events.

Next, I use this modified empirical model to derive the estimating equations in Section V. Consider first the queueing efficiency estimating equation, for which I must derive an expression for (expected) loading time L_{ij}. After I observe a bus arriving in the loading bay, two shocks must occur for a bus to depart: first, the newly-introduced bus loading shock, and, second, the bus departure shock, which, recall, occurs at Poisson rate $\mu p_{ij}/\overline{\eta}$. Because the two Poisson shocks are independent, this measured loading time, in expectation, equals the mean of the corresponding hypoexponential distribution, i.e.

$$L_{ij} = \frac{1}{\iota_{ij}} + \frac{1}{\mu p_{ij}/\overline{\eta}} = \frac{\overline{\eta}}{\mu p_{ij}} + \underbrace{\frac{1}{\iota_{ij}}}_{\equiv \varepsilon_{ij}}, \tag{C.6}$$

where I define the idiosyncratic error $\varepsilon_{ij} \equiv \frac{1}{\iota_{ij}}$. Rewriting this equation at the individual bus departure level yields my main specification (13) for μ. In turn, for bus arrival efficiency ρ, my specification (14) is simply a bus-departure-level version of (C.5).

Table D.1. Queueing Efficiency: First Stage

	(1) p_{ijtd}^{-1}	(2) p_{ijtd}^{-1}	(3) p_{ijtd}^{-1}
log number commuters	-0.26	-0.38	-0.49
	(0.05)	(0.05)	(0.14)
Route-by-Date FE		✓	✓
Origin-by-Time-by-Date FE			✓
Observations	1077	1077	1023
F Statistic	26.05	69.37	12.39

Notes: Robust standard errors in parentheses, clustered at the route-by-five minute time period-by-date level. Table presents the first-stage specifications from 2SLS estimation of (13) over bus departures in my station count data. The first stage regresses the (inverse) probability of a nonzero passenger queue conditional on a bus being present in the loading area in the rolling hour centered around a bus's departure on the log number of commuters living in the transport analysis zone (TAZ) where the route originates and working in the route's destination TAZ who report leaving their home during this same rolling hour, calculated from the 2013 Cape Town Household Travel Survey, along with fixed effects, as noted, for route-by-date and origin-by-five minute time period-by-date. I report the effective first-stage F statistic following Olea and Pflueger (2013).

D. Estimation

D.1 Queueing Efficiency

D.1.1 First Stage Specifications

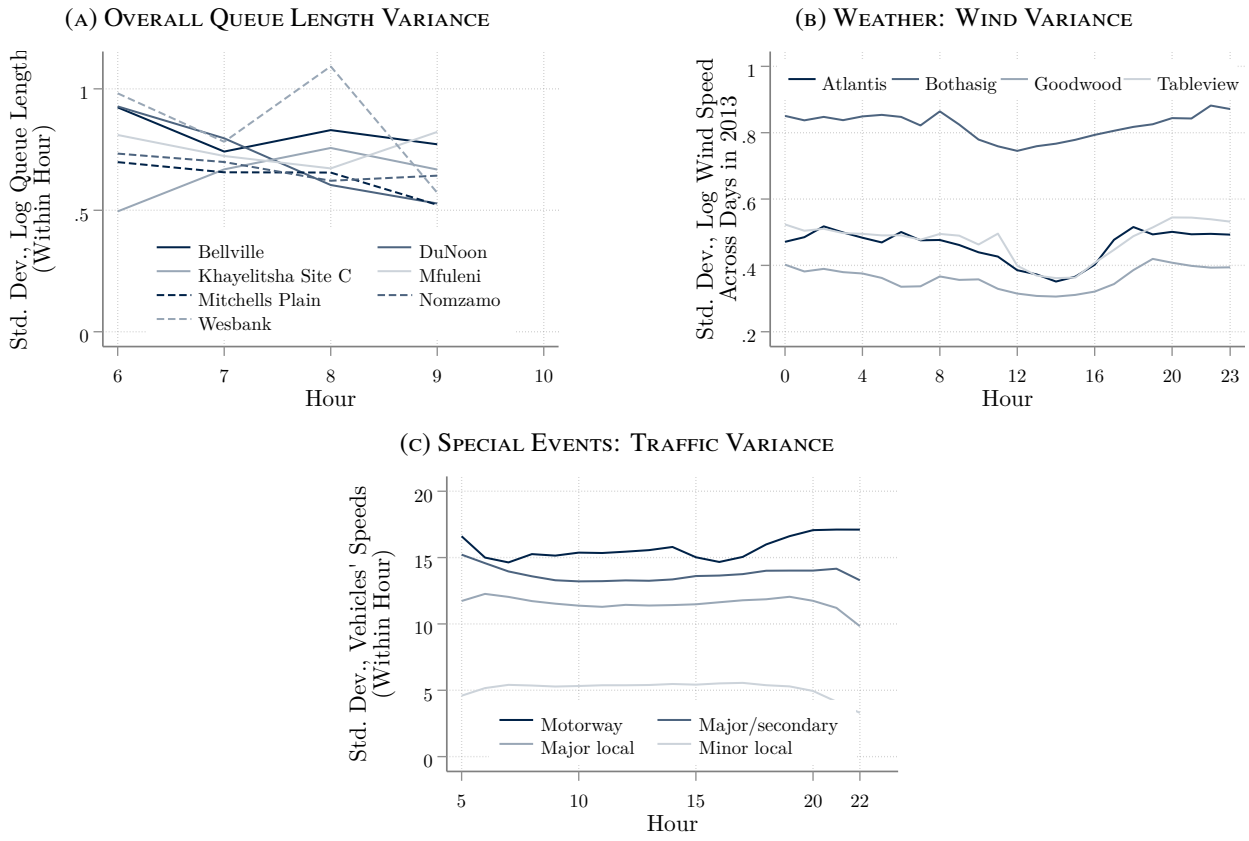
In this section, I present the first-stage regressions from 2SLS estimation of Equation (13) in Table D.1. Table 2 in the main text, in turn, contains the second-stage results.

D.1.2 Threats to Identification

I now probe the plausibility of the exclusion restriction for the queueing efficiency regression: namely, that idiosyncratic interruptions to loading do not vary systematically with work start times. Recall that the nine-year lag between the data collection of my instrument and the rest of the variables means that only very persistent (idiosyncratic) loading interruptions could threaten identification. Nonetheless, some sources of delay might indeed change little over time – and happen to recur at times when more people start work. Most sources of loading interruptions are likely short-lived and would thus increase the variance of queue lengths at the affected station within a given hour; in Figure D.1a, I therefore examine exactly this within-hour queue standard deviation in my station count data over different hours, separately for each origin station. While stations differ in the variability of their queues, this within-hour variance does not seem to change systematically by hour of the morning commute, a first hint that loading disruptions seem unlikely to be correlated with work start times.

Next, I consider specific sources of interruptions. In Figure D.1b, I employ wind speed as an indicator of bad weather and plot, for four measuring stations in Cape Town, the standard deviation of hourly log wind speed, calculated across days within the year. The fact that the day-by-day variability in wind speed does not differ by morning commute hour suggests that bad weather does not occur disproportionately often at certain work

FIGURE D.1. QUEUEING INTERRUPTIONS: DIFFER BY HOUR?



Notes: Panel (A) displays the standard deviation of log queue lengths across 5-min. periods and routes within a given hour, separately for each origin station, from my rank count data. Panel (B) plots the standard deviation of log wind speed across days within 2013 by measuring station in Cape Town; data obtained from the City of Cape Town Open Data Portal. Panel (C) displays the standard deviation of vehicle speeds on a given street segment during a given hour on a sample date in 2021, averaged over streets within an aggregated functional road class. Speed data comes from the TomTom MOVE Traffic Stats API.

start times. Finally, in Figure D.1c, I use on-road traffic as an indicator for special events, road closures, and the like, which might disrupt the loading process. Specifically, I plot the standard deviation of speeds across vehicles on the road in Cape Town at different hours, averaged for various classes of roads. Logically, variance tends to fall at night for most road classes, and motorways have a higher variance of vehicle speeds. However, to the extent that special events or closures cause variation in speeds across vehicles within a given hour, the frequency of events would not seem to change throughout the 6-10am morning commute. In sum, available evidence suggests that key sources of minibus loading delays vary little over the course of the morning rush hour, so any correlation with my work start times instrument might likely be mild.

D.1.3 Alternative Instruments and Estimation Strategies

I now seek to verify the robustness of my estimated queueing efficiency. First, I test several alternative instruments and re-estimate versions of Equation (13); Online Appendix Table D.2 displays the results. In

Column 1, I interact the baseline instrument, the number of commuters along route ij leaving home within the rolling hour around a bus's departure, with an indicator for origin stations i located within 1 kilometer of MyCiti BRT trunk routes. My baseline instrument is measured in 2013, and most of these BRT trunk routes opened soon after, so past commuter numbers might have a larger effect on current queue prevalence for origins that did not gain BRT access in the meantime. While the first stage is less strong than in the baseline specification in Table 2, I find a similar $\hat{\mu}$ whose 95% confidence interval includes my baseline estimate.

In Column 2, I instead use as an instrument the *index of dispersion* of commuters' times leaving home, calculated as the variance-to-mean ratio of the number of commuters leaving home across 5-minute intervals within the same rolling hour as for the baseline IV. A greater variance in work start times will go hand-in-hand with the larger passenger arrival shocks which lead to the build-up of queues and thus should substantially affect queue prevalence on the right-hand side of Equation (13). The advantage of using a variance-based measure of demand is that I can then, in Column 3, control for the *level* of commuter demand, i.e. the total number of commuters leaving home in the rolling hour window. Certainly, this variation, measured on the same day as the loading times and queue prevalence, would be highly correlated with weather shocks. However, my index of dispersion is effectively lagged by 9 years. Conditional on the level of commuter demand, any persistent variation in times leaving home within a given hour might be even less likely to be correlated with the time-varying weather and special event shocks 9 years later in the error term. When controlling for the level of commuter demand, in Column 3, I find a somewhat higher queueing efficiency, relative to my baseline estimate of $\hat{\mu} = 3.54$, though the original estimate remains within the corresponding 95% confidence interval. Nevertheless, for further reassurance, I later test the robustness of my central counterfactual to higher queueing efficiency.

Second, I pursue an alternative, simpler estimation strategy for μ that does not rely on estimates of queue prevalence and might be less susceptible to bias from loading interruptions due to weather or special events. In particular, I must assume that loading interruptions become negligible when queue length n_{ijtd} in passengers equals or exceeds a full busload of $\overline{\eta}$, perhaps because drivers and passengers become more hurried in the presence of long queues and no longer allow weather or other crowds to cause delays. Through the lens of my model, then, the expected loading time for buses which arrive on route ij at time t on date d to find $n_{ijtd} > \overline{\eta}$ satisfies $E\left(L_{ijtd} | n_{ijtd} \geq \overline{\eta}\right) = \frac{\overline{\eta}}{\mu}$, since the first $\overline{\eta}$ passengers waiting will immediately board when they reach the front of the queue. I operationalize this strategy by calculating the average loading time for the first bus which arrives within each time period t, where, at the start of the period, I observe at least $\overline{\eta}$ passengers in the queue. The implied queueing efficiency, $\hat{\mu} = 4.46$, as displayed in Online Appendix Table D.3, remains qualitatively similar to the baseline. Nonetheless, the new 95% confidence interval, at $[3.74, 5.19]$, marginally fails to include my baseline estimate of $\hat{\mu} = 3.54$, which further motivates the aforementioned sensitivity analysis with respect to higher queueing efficiency values.

D.2 Minibus Arrival Efficiency: Details

As discussed in Section V, I estimate a model-implied equation for the expectation of the gap, gap_{sijhd}, in minutes between the departure of the previous bus from the origin station and the arrival of a given bus s on

Table D.2. Queueing Efficiency: Alternative Instruments

	IV:		
	(1) Log(# Commuters) × BRT	(2) Log(Commuter Index of Dispersion)	(3) Log(Commuter Index of Dispersion)
μ	4.01	3.61	4.51
Queueing Efficiency	(1.77)	(1.90)	(2.25)
# Commuters Control			✓
Route-by-Date FE	✓	✓	✓
Origin-by-5 Min.-by-Date FE	✓		
Origin-by-Hour-by-Date FE		✓	✓
Observations	1023	1005	1005
First-Stage F Statistic	6.48	6.18	8.49

Notes: Robust standard errors in parentheses, clustered at the route-by-five minute time period-by-date level. Table presents estimates of (13) over bus departures in my station count data, with fixed effects included, as noted, for route-by-date, origin-by-five minute time period-by-date, or origin-by-hour-by-date. In Column 1, I instrument for queue prevalence, i.e. the probability of a nonzero passenger queue conditional on a bus being present in the loading area in the rolling hour centered around a bus's departure, using the log number of commuters living in the transport analysis zone (TAZ) where the route originates and working in the route's destination TAZ who report leaving their home during this same rolling hour, interacted with a dummy variable for routes whose origin rank is located within 1km of a MyCiti BRT trunk route. In Column 2, I instead instrument using an index of dispersion of work start times, calculated as the variance-to-mean ratio of the number of commuters leaving home during each 5-minute interval within the rolling hour around a bus's departure. In Column 3, I additionally include the total number of commuters leaving home in this rolling hour window itself as a control variable. I additionally report the effective first-stage F statistic following Olea and Pflueger (2013). Finally, I obtain the implied value of μ by taking the reciprocal of the coefficient on (inverse) queue prevalence and multiplying by the externally-calibrated minibus capacity $\overline{\eta}$.

route ij from neighborhood (Transport Analysis Zone, i.e. TAZ) $n(i)$ to neighborhood $n(j)$ during hour h on date d in my station count data:

$$\text{gap}_{sijhd} = \frac{1}{\rho}\log\left\{\exp\left[\rho\left(\frac{2T_{ij}}{b_{ijhd}} - L_{sijhd}\right)\right] + 1\right\} + \underbrace{\Gamma_{n(i)hd} + \Gamma_{n(j)} + \vartheta_{sijhd}}_{\equiv -\frac{1}{\rho}\log \varsigma_{sijhd}}. \tag{D.1}$$

The unobserved arrivals efficiency, ς_{sijhd}, reflects a variety of factors outside the model likely to impact associations' bus supply: overall foot and car traffic levels in origin neighborhoods as well as weather, special events, and road closures. I neutralize traffic and weather around the origin minibus station with origin neighborhood-by-hour-by-date fixed effects and delays common to destinations with destination neighborhood fixed effects. To address any differential impacts of, say, foot traffic on different routes operating out of the same origin neighborhood, I instrument for the first term on the right-hand side, which varies primarily with bus supply b_{ijhd} on a route, with the distance from that route's origin to destination. Route length determines operations costs and thus supply but should not systematically vary with the degree to which, say, local traffic or road closures affect specific routes at the same station.

In my baseline specification in Column (1) of Table D.4, I find a bus arrival efficiency of $\hat{\rho} = \texttt{0.177}$. When I

Table D.3. Queueing Efficiency via Mean Loading Time

Parameter	(1) $E\left(L_{ijtd}\vert n_{ijtd} \geq \overline{\eta}\right)$
μ	4.46
Queueing Efficiency	(0.37)
Observations	177

Notes: Standard error in parentheses. Table presents estimates of queueing efficiency using the relation $E\left(L_{ijtd}\vert n_{ijtd} \geq \overline{\eta}\right) = \frac{\overline{\eta}}{\mu}$ in my station count data, where I take the mean loading time of the first bus to depart during any five-minute period which starts with at least $\overline{\eta} = 15$ waiting in the queue for the corresponding route. I obtain the implied value of μ by taking the reciprocal of this expectation and multiplying by $\overline{\eta}$.

Table D.4. Minibus Arrival Efficiency Estimates

	GMM-IV	
	(1) bus gap	(2) bus gap
ρ	0.177	0.178
Arrival Efficiency	(0.09)	(0.09)
Origin Neighborhood-by-Date FE	✓	
Destination Neighborhood FE	✓	✓
Hour FE	✓	
Origin Neighborhood-by-Hour-by-Date FE		✓
Observations	1,158	1,158

Notes: Robust standard errors in parentheses. Table presents estimates of (D.1) over bus departures in my station count data, with fixed effects included, as noted, and neighborhoods are the 18 transport analysis zones (TAZ) in Cape Town. Consistent with my model, I restrict the mean of each set of fixed effects to equal zero. I estimate each specification by two-step GMM with one instrument: the log straight-line distance from a route's origin to destination.

add origin neighborhood-by-hour-by-date fixed effects in Column (2), the estimate remains almost unchanged; I employ the Column (2) estimate in the model quantification and counterfactuals in Sections VI-VII.

D.3 Stated Preferences

D.3.1 Sample Representativeness

I noted in Online Appendix B.2.6 that my new stated preference survey oversamples minibus users. I test for bias resulting from this non-representative sample in three ways. First, I re-estimate the model using the city-conducted survey plus only those respondents in my survey interviewed at a non-minibus station location – namely the Middestad Mall/Bellville intermodal interchange – a recruitment location less prone to oversampling of minibus riders. Column 2 of Online Appendix Table D.5 shows that the estimated parameters, though somewhat noisier, mirror my full-sample estimates quite closely, in particular the rate of time preference r, Gumbel scale ν, and the high value that high-skill commuters place on minibus station security. Even this non-minibus station sample, however, still oversamples minibus commuters. Second, in Column 4, I thus weight my own sample, which, critically, does not contribute to the identification of the

Table D.5. Stated Preference: Robustness to Sample

Parameter	Skill	(1) Baseline	(2) Non-Minibus Station + City-Run Sample Only	(3) Minibus Station + City-Run Sample Only	(4) Commute Mode-Weighted
r		0.001	0.0014	0.0013	0.0011
		(0.0004)	(0.0007)	(0.0006)	(.0005)
ν		4.76	6.81	6.47	5.83
		(1.26)	(2.72)	(2.38)	(1.98)
κ_M	*Low*	7.68	10.59	10.11	9.24
		(1.56)	(3.53)	(3.06)	(2.55)
	High	15.03	21.11	20.10	18.2
		(3.55)	(7.79)	(6.78)	(5.66)
ξ_{security}	*Low*	-1.09	-2.12	-1.20	-1.55
		(0.39)	(1.06)	(0.62)	(0.69)
	High	-2.75	-4.90	-3.44	-5.1
		(0.84)	(2.29)	(1.42)	(1.85)
$\xi_{\text{no overloading}}$	*Low*	-1.38	-2.02	-1.78	-1.26
		(0.437)	(1.01)	(0.77)	(0.595)
	High	-1.39	-1.25	-2.01	-1.43
		(0.543)	(1.28)	(0.97)	(0.83)
$\xi_{\text{no speeding}}$	*Low*	-1.36	-3.03	-1.27	-2.12
		(0.44)	(1.37)	(0.64)	(0.85)
	High	-0.825	-1.86	-0.93	-0.582
		(0.465)	(1.39)	(0.73)	(0.73)
κ_P	*Low*	3.63	4.52	4.38	4.14
		(0.51)	(1.08)	(0.94)	(0.80)
	High	9.17	12.5	11.94	10.95
		(1.89)	(4.18)	(3.64)	(3.04)
N Respondents		820	546	681	820

Notes: Robust standard errors in parentheses. The unit of analysis is an alternative by choice set by individual respondent in either my newly collected minibus stated preference survey (in Cape Town, estimates reflect $N = 489$ unique individuals) or a stated preference module of the 2013 Cape Town Household Travel Survey ($N = 646$ unique individuals). The estimated parameters are derived from the coefficients in a multinomial logit model with choice probabilities given by (15). Column 1 displays the baseline estimates, as in Table 3; Column 2 estimates the model on only the 2013 city-run survey respondents plus the respondents in my survey interviewed at a non-minibus station location, i.e. the Middestad Mall/Bellville transport interchange. Column 3 estimates the model on only the city-run survey respondents plus the respondents in my survey interviewed at minibus stations. Finally, Column 4 estimates the model on the full sample but weights the respondents in my survey by the aggregate citywide share of their reported commute mode divided by that mode's share among respondents to my survey.

relative utility costs across modes, by the ratio between the citywide mode share, from the 2013 Household Travel Survey, and the in-sample mode share of a respondent's self-reported commute mode. Reassuringly, the key takeaways remain.

Third, I use only my own stated preference sample to re-estimate a version of the logit model corresponding to Equation (15) without mode-specific utility costs, since my survey does not include variation in transport modes. I use the resulting Gumbel scale parameter ν, empirical aggregate mode shares, and prices – the median empirical minibus fare, median calibrated public fare, and calibrated car per-commute cost – to compute model-implied own-price demand elasticities for each mode. Table D.6 displays the results, side-by-side with

Table D.6. Stated Preference: Demand Elasticities by Sample

Mode	(1) Baseline	(2) My Survey Only
Minibus	0.17	0.19
Public Bus/Train	0.17	0.19
Car	1.62	1.83

Notes: This table displays own-price elasticities of demand implied by two different estimates of the Gumbel scale parameter *nu*. In Column (1), the Gumbel scale parameter is estimated from the full stated preference sample, as in Table 3, and in Column (2), I estimate an equivalent model without mode-specific utility costs using only my newly collected minibus stated preference survey. To calculate demand elasticities, I additionally use (i) median minibus fares from my on-board tracking data as well as the median calibrated public fare and calibrated car per-commute cost and (ii) aggregate mode shares from the 2013 Cape Town Household Travel Survey.

Table D.7. Stated Preference: Value of Security by Interview Location

	$\lvert\xi_{\text{security}}\rvert$		
Skill	Bellville/Middestad Mall	Khayelitsha Site C	Somerset West Shoprite
Low	-1.59	-1.27	-0.23
	(0.60)	(0.56)	(0.50)
High	-3.46	-3.42	-1.77
	(1.24)	(1.16)	(0.80)

Notes: Robust standard errors in parentheses. Table displays values of security, ξ_{security}, by skill group and stated preference interview location, estimated from a multinomial logit equivalent to (15), except with interactions between interview location dummies and the skill-specific security effects. I estimate all effects in one specification across alternatives, choice sets, and individuals in my own and the 2013 Cape Town Household Travel Survey stated preference modules.

the demand elasticities implied by the baseline ν from Table 3. Using only my own stated preference sample yields a system of demand elasticities quite close to that from the full sample, providing reassurance that my oversampling of minibus users does not substantially alter key model-implied demand patterns.

D.3.2 Selection by Interview Location

I now examine whether respondents' interview location in my stated preference survey affects the estimated parameters. Recall that in Column 2 of Online Appendix Table D.5, I re-estimate the model using only those respondents from my own survey interviewed at non-minibus station locations, while in Column 3, I use only those interviewed at minibus stations; in both cases, estimation also uses the representative sample from the city-run survey. The estimates remain broadly similar in either subsample, though the non-minibus station sample yields somewhat higher values of security and banning speeding. Thus, interviewees present in minibus stations would not seem to be strongly selected on their cost sensitivity, value of time, or, e.g. preferences for minibuses relative to cars.

To more precisely quantify the aforementioned differences in security value by interview location, I re-estimate my stated preference model on the full sample, but now including interactions between interview location dummies and the skill-specific security effects. Online Appendix Table D.7 displays the resulting values of

security by interview location and skill group; regardless of interview location, the high-skill more strongly value security. The security values for non-minibus-station respondents in Bellville and for in-minibus-station respondents in Khayelitsha are similar and relatively precisely estimated, providing some reassurance that minibus station users do not display a consistently lower preference for security. Nonetheless, the noisier, lower values of security for respondents at the Somerset West minibus station do point in the direction of minibus station respondents being somewhat selected toward lower security valuations. In that case, my estimated security values might in fact understate the average value of security among the general population.

D.3.3 Alternative Revealed Preference Estimation Strategy

For comparison, I use the data on commuters' mode choices from the 2013 Cape Town Household Travel Survey to estimate a multinomial logit mode choice model by more-standard revealed preference methods. Note that the commuter survey data lacks information on quality improvements and does not separately enumerate wait and travel time. I thus estimate a modified version of Equation (15) over commuters i and mode alternatives l,

$$\pi_{il}^{g} = \frac{\exp\left(-\kappa_{m(l)}^{g} - r\omega_i C_l - \tau_l\right)^{1/\nu}}{\sum_{l'} \exp\left(-\kappa_{m(l')}^{g} - r\omega_i C_{l'} - \tau_{l'}\right)^{1/\nu}}, \tag{D.2}$$

where I now let C_l denote total commute time for alternative l, $\kappa_{m(l)}^{g}$ continues to denote mode m-specific utility costs by skill group g, τ_l denotes the monetary cost of alternative l, and ω_i denotes respondent i's personal income. Using the household travel survey data, I calculate commute time and monetary costs as origin-by-destination transport analysis zone (TAZ) averages for each mode, and personal income ω_l equals household income per employed member. I again normalize $\kappa_A^g = 0$ for the car alternative A and restrict the sample to individuals employed outside the home between 25 and 65 years of age. Because the survey, unlike my model, includes within-TAZ commutes and thus a nontrivial share of non-motorized commutes, I also include a fourth walk/bike mode.

The coefficient on fares, which would identify the Gumbel scale ν, turns out to be estimated very imprecisely, perhaps due to unobservable quality-related dimensions of different modes. Indeed, such omitted variables are one of the primary reasons I employ stated rather than revealed preferences in my main estimation procedure. I thus instead fix ν to equal the stated preference estimate of `4.76` and re-estimate only the commuter rate of time preference r and the mode-specific utility costs κ_m^g. Recall that, since the revealed preference data lack information on quality improvements, I cannot estimate values of security or the other specific quality-related attributes.

Online Appendix Table D.8 displays the resulting parameter estimates. The revealed preference estimates yield a slightly higher commuter rate of time preference, $r =$ `0.002`, though the 95% confidence interval of [`0.0012`, `0.0028`] comes close to including my baseline stated preference estimate. The revealed preference utility costs for public bus/train and minibuses are negative for low-skill commuters, indicating, in contrast to the stated preference results, that the low-skill actually prefer these shared modes to cars. This less plausible

Table D.8. Revealed Preference Demand Estimates

Parameter	Value	
r	.002	
Commuter Rate of Time Pref.	(.0004)	
Utility Costs	*Low-Skill*	*High-Skill*
κ_P	-0.65	7.90
Public Bus/Train Utility Cost	(0.18)	(0.47)
κ_M	-0.55	8.58
Minibus Utility Cost	(0.21)	(0.53)

Notes: Robust standard errors in parentheses. Estimates reflect $N = 34,573$ individuals by mode alternatives, or $8,785$ unique commuters, in the revealed preference household survey component of the 2013 Cape Town Household Travel Survey, where commuters reported their actual commute modes. The estimated parameters come from a multinomial logit model with choice probabilities given by (D.2). I normalize $\kappa_A^g = 0$ and restrict the sample to individuals employed outside the home between 25 and 65 years of age. I calculate commute times and monetary costs as origin-by-destination transport analysis zone (TAZ) averages for each mode. Note that I fix the Gumbel scale parameter ν to equal the stated preference estimate of `4.76` and re-estimate only the commuter rate of time preference r and the mode-specific utility costs κ_m^g.

result could again be interpreted as pointing towards omitted variables bias. High-skill commuters' utility costs, on the other hand, qualitatively align with the stated preference estimates, in that they strongly disfavor both minibuses and public bus/train relative to cars.

D.4 Externally Calibrated Parameters

D.4.1 Geography

The model geography consists of the $I = 18$ transport analysis zones (TAZ) in Cape Town. I exclude from commuters' home and work choice sets one location with fewer than 5,000 residents employed within Cape Town and in which fewer than 5,000 Cape Town residents work. I also exclude the few home-work location tuples with no existing commuters. Finally, since my model is not suited to short-distance commutes, I do not allow commuters to choose to live and work in the same location.

D.4.2 Commuter Populations N^g and Average Wages

I calibrate commuter populations N^g using the 2013 Cape Town Household Travel Survey (see Online Appendix B.3) and the accompanying sample weights. Since commuters in my model cannot work in their home zone, I exclude those who work in the same transport analysis zone, as well as the less than 5,000 residents of the aforementioned one primarily non-residential TAZ, from these commuter populations. Since my model and data apply to the 6-10am peak-hour commute, I rescale these populations by the share (84%) of Cape Town commuters who start work within these four hours and then divide by 240 to calculate N^g as a per-minute inflow. I define two skill groups g, high and low, where high-skill includes those with a tertiary degree. For the stated preference estimation and normalization of model wages, I also use the 2013 Cape Town survey to compute average wages by skill group, taking the weighted mean daily per-person household

Table D.9. Estimation of Road Congestion Elasticity γ

Variable	(1) log mean travel time	(2) log mean travel time
log traffic volume	0.0917	
	(0.000451)	
Effect of log traffic volume *for...*		
Motorway		0.0244
		(0.00106)
Major/Secondary		0.0977
		(0.00106)
Major Local		0.0968
		(0.000877)
Minor Local		0.0898
		(0.000613)
Segment FE	✓	✓
Observations	2,355,901	2,355,901
R-Squared	0.023	0.023
Number of Segments	231,707	231,707

Notes: Robust standard errors in parentheses. The unit of analysis is a road segment in Cape Town ($n = 231,707$) by hour block of a sample weekday. Each column presents the estimated congestion elasticity derived from a regression of the log average travel time over the segment on the log traffic volume, both from the TomTom MOVE Traffic Stats API. Column 2 interacts traffic volume with four aggregated categories of functional road class. Motorway designates functional road class 0, major/secondary classes 1-3, major local classes 4-5, and minor classes 6-7. Both specifications include segment fixed effects.

income of workers in a skill group employed outside the home.[63]

D.4.3 Road Congestion Elasticity γ

Table D.9 displays the results from a regression, in TomTom MOVE API data for Cape Town, of log segment travel time on traffic volume, additionally interacted with road type indicators. I further discuss the specification in the main text in Section V.

D.4.4 Free-Flow Driving Times $\bar{t}_{ik}$

I calibrate the free-flow travel time $\bar{t}_{ik}$ between (centers of employment of) TAZ i and k using the Google Maps Distance Matrix API. To approximate the no-traffic travel times, I query predicted travel times for Sunday, May 8, 2022 at 11pm. The driving distance that determines minibuses' operating costs equals the distance driven along the route chosen in these queries.

[63] To calculate daily personal income ω_i in the Cape Town survey, I take the midpoint of the household's income bin, divided by 22.5 (the number of working days in a month) times the number of people in the household. Additionally, I multiply by $\frac{1}{2}$ since I only model a one-way commute. For my own survey, I make similar adjustments, except that I have personal income directly instead of needing to impute it from household income. Finally, I convert all monetary amounts, including fares, to USD for scaling purposes.

D.4.5 Public Bus/Train Wait and Travel Times H_{ij}, T_{ijP}

I calibrate the full origin-destination matrix of public bus/train wait, H_{ij}, and travel times, T_{ijP}, using a stylized network. The public network consists of links between every pair of locations, and the Microsoft Azure API provides average public bus/train wait and travel times along each link, where, crucially, these calculations allow commuters to walk to transit stops and transfer, if necessary.[64] Then, for each origin-destination pair ij, I find the shortest path through the network, accounting for wait and travel time along each link; the respective sums yield the total wait time, H_{ij}, and the total travel time, T_{ijP}, even for origin-destination pairs where no direct connection exists.

D.4.6 Public Bus/Train Fares τ_{ijP} and Car Commute Cost τ_A

Public bus/train fares for a given route ij are calculated using the Cape Town MyCiti bus rapid transit distance-based fare scheme, where I make the calculation using the straight-line distance between TAZ centroids. I calculate the car monetary commute cost τ_A using an "average [monthly] total mobility cost " calculated by WesBank of South Africa.[65]

D.4.7 Minibus Operations Parameters $\overline{\eta}$, χ_{ij}

I set minibus capacity to the $\overline{\eta} = 15$ passengers in line with the size of 94% of minibuses, as discussed in Fact 3. I parameterize the minibus operating cost matrix χ_{ij} as proportional to driving distance and then use fuel efficiency figures provided by the firm GoMetro to calibrate the per-kilometer cost.[66]

E. Validation and Results

E.1 Mode Shares and Fares, Model vs. Data

In this section, I compare equilibrium quantities predicted by my model to those observed in the data. First, in Online Appendix Table E.1, I show that, for both minibus and car, the model-predicted mode shares by origin-destination pair and skill are significantly positively correlated with their empirical counterparts. In a similar vein to the validation tests conducted by Tsivanidis (2023), the fact that minibus mode share slope is statistically indistinguishable from 1 shows that my model, on average, replicates the spatial variation in mode shares, idiosyncratic factors outside the model notwithstanding.

Second, in Online Appendix Table E.2, I show that the model-predicted route-level minibus fares predict the

[64] Specifically, from the Microsoft Azure API (similar to Google Maps), I obtain wait and travel time for each link in the public network from averages over 6 evenly-spaced trips on a Wednesday between 7:00 and 8:00am via public modes: Metrorail commuter trains, Golden Arrow private scheduled buses, and MyCiti bus rapid transit. The public bus/train wait time equals the time between the queried departure time and the actual departure time of the suggested itinerary. Travel time includes any walking and transfer time.

[65] The monthly average total mobility cost equals ZAR 9,356.80; I divide this figure by the number of (half-)working days per month, and convert to USD.

[66] I use the driving distance under free-flow (Sunday, 11pm) conditions between the centers of employment of transport analysis zones (TAZ) i and j predicted by the Google Maps Distance Matrix API. For the per-kilometer cost, I multiply the Toyota Quantum minibus's litres of diesel used per kilometer, 0.099, by the June 2022 diesel per-litre price in South Africa, ZAR22.63, and convert to USD, in line with other prices in my model.

Table E.1. Mode Choice Probabilities, Data vs. Model

	Minibus	Car
Variables	Mode Share, Data	Mode Share, Data
Mode Share, Model	1.232	1.182
	(0.156)	(0.0944)
Constant	-0.133	0.271
	(0.0362)	(0.0299)
Observations	483	483
R-squared	0.105	0.247

Notes: This table presents the results of OLS regressions of empirical mode shares on those predicted by the model at the skill group by origin by destination transport analysis zone level. In the data, I calculate skill-specific origin-destination transport analysis zone (TAZ) commute mode shares from the 2013 Cape Town Household Travel Survey, taking shares of respondents working outside the home who commute by each mode. In the model, I take the share of commuters with a given skill and chosen home and work location (TAZ) who use a given mode.

Table E.2. Minibus Fares, Data vs. Model

	Minibus
Variables	Fare, Data
Fare, Model	1.243
	(0.515)
Constant	-0.0413
	(0.520)
Observations	22
R-squared	0.207

Notes: This table presents the results of OLS regressions of mean minibus fares in my station counts at the origin by destination transport analysis zone level on those predicted for the corresponding route by my model. Note that, in the model, origin-destination TAZ pairs are synonymous with routes.

corresponding average fares by origin-destination transport analysis zone pair in the data. Though somewhat noisier, the coefficient's 95% confidence interval continues to include 1.

E.2 Stated Preference Heterogeneity

In this section, I estimate heterogeneity in demand parameters by a series of demographic characteristics and find plausible heterogeneity in preferences. I take Equation (15) and interact a dummy variable for the binary demographic characteristics listed in Column 1 of Online Appendix Table E.3 with the terms in the multinomial logit model that identify utility costs, the value of quality improvements, and the value of time. Each column of Online Appendix Table E.3 displays the coefficient on the interaction of a demographic characteristic with a given parameter; for ease of interpretation, in the case of the quality improvement valuations, I report the effects of each demographic characteristic on the *absolute value* of the attribute's associated reduction in utility costs. In other words, a negative coefficient indicates that a quality improvement reduces utility costs by *less* for a given group, such that the group places a lower value on that improvement. I find that women and college workers have a higher value of time saved, even conditional on income; the

Table E.3. Stated Preference Heterogeneity: Difference in Parameter Estimate, Versus Base Category

		Mode Utility Cost		*Effects on Minibus Utility Cost*		
Dimension	r	κ_M	κ_P	$\lvert\xi_{\text{overload}}\rvert$	$\lvert\xi_{\text{security}}\rvert$	$\lvert\xi_{\text{speed}}\rvert$
Female	0.0013	-3.61	-3.27	-0.222	-1.33	-0.49
	(0.0006)	(1.06)	(0.924)	(0.419)	(0.535)	(0.436)
College	0.0019	6.66	4.62	0.052	1.71	-0.458
	(0.0007)	(1.94)	(1.28)	(0.481)	(0.659)	(0.499)
Age>45	0.0027	-1.03	-1.80	0.494	1.72	2.50
	(0.001)	(0.709)	(0.671)	(0.640)	(0.770)	(0.906)

Notes: Robust standard errors in parentheses. Each cell gives the coefficient on the interaction of a dummy variable for the demographic characteristic listed in Column 1 with the parameter at the top of each column in a multinomial logit equivalent to (15). In the final three columns, I report the effect of each demographic characteristic on the absolute value of the utility cost reduction associated with a given quality improvement. I estimate all interaction effects in each row in one specification across alternatives, choice sets, and individuals in my own and the 2013 Cape Town Household Travel Survey stated preference modules.

former result echoes Borghorst et al. (2021), who find that women's marginal cost of commuting increases after the birth of children. That college-educated workers value their time more highly, even conditional on income, might similarly reflect a higher value of home production. Surprisingly, women place a lower value on security; perhaps men are more likely to be involved in gang activities that would put them at risk. Older workers place a higher value on security and especially on driver adherence to speed limits, suggesting an intuitive greater risk aversion.

E.3 Counterfactual Robustness

In this section, I show how alternative parameter values or modeling assumptions alter the welfare gains from the optimization of minibus demand and supply, as discussed for my baseline model in Section VII. In each case, I re-calibrate amenities and the (exogenous component of) wages as well as the association bargaining power β and driver wage $\overline{\omega}$, as in Section V. I then re-solve for the baseline equilibrium as well as the social planner optimum under each alternative parameter value or modeling assumption.

Higher Queueing Efficiency μ

First, I impose a 25% higher queueing efficiency, $\hat{\mu} = 4.42$. The new internally-calibrated bargaining power $\hat{\beta}$ remains similar to the baseline, but the driver wage $\overline{\omega} = 0.0036$ comes in higher. I then re-solve for the social planner optimum; the first line of Online Appendix Table E.4 lists the planner-induced changes in aggregate mode shares and the welfare gains, relative to the status quo, all given the new higher value of μ. The planner would increase minibus mode shares and decrease car mode shares by similar magnitudes to the baseline, and the welfare gains remain similar, as they stem largely from the fundamental Market Mohring Effect and filling buses of a given size with greater passenger demand.

Higher Bus Arrival Efficiency ρ

Second, I impose a 25% bus arrival efficiency, $\hat{\rho} = 0.22$. As in the case of higher queueing efficiency, after internal re-calibration, association bargaining power $\hat{\beta}$ remains similar to the baseline, and the driver wage increases, albeit more mildly, to $\overline{\omega} = 0.0012$ The second line of Online Appendix Table E.4 characterizes the social optimum given the new higher value of ρ; changes in mode shares and welfare again mirror those under the baseline calibration.

Nested Logit

Third, I allow for differential substitution patterns over locations versus over modes – for which the quantitative spatial literature provides ample evidence (Ahlfeldt et al. (2015) and Tsivanidis (2023)). Specifically, I assume a nested logit demand structure. Commuters draw an idiosyncratic preference ε_{ij} for each home-work location pair, with variance scaled by the parameter ζ, as well as, for each home, work, and mode combination, an idiosyncratic mode preference shock $\varepsilon_{m|ij}$ with variance scaled by the same parameter ν as in the main text. Commuter utility then reads $\theta_i^g + U_{ijm}^g + \omega_j^g + \zeta\varepsilon_{ij} + \nu\varepsilon_{m|ij}$. I solve commuters' problem by backward induction; having chosen a home location i and work location j, commuters' mode choice probabilities follow

$$\pi_{m|ij}^g \equiv \frac{\exp\left(U_{ijm}^g\right)^{1/\nu}}{\sum_{m'}\exp\left(U_{ijm'}^g\right)^{1/\nu}}. \tag{E.1}$$

I define $\overline{U}_{ij}^g \equiv \nu \log\left[\sum_{m'}\exp\left(U_{ijm'}^g\right)^{1/\nu}\right]$ such that commuters then choose a given home-work pair with probability

$$\pi_{ij}^g \equiv \frac{\exp\left(\theta_i^g + \overline{U}_{ij}^g + \omega_j^g\right)^{1/\zeta}}{\sum_{i',j'}\exp\left(\theta_{i'}^g + \overline{U}_{i'j'}^g + \omega_{j'}^g\right)^{1/\zeta}}. \tag{E.2}$$

The home-work-mode location choice probabilities then satisfy

$$\pi_{ijm}^g \equiv \pi_{ij}^g \pi_{m|ij}^g. \tag{E.3}$$

Equations (E.1)-(E.3) then replace Equation (8) in the equilibrium definition and pin down choice probabilities π_{ijm}^g.

I set $\nu = 4.76$ as in the main text and, given the lack of available estimates for South Africa, calibrate ζ by appropriately transforming one of the few available estimates from a developing-country context, namely a Frechet elasticity of location choice from Tsivanidis (2023). He estimates a home and a work location choice elasticity for low- and high-skill workers separately. As a form of upper bound on the (un)willingness of commuters to change home or work locations, I appropriately transform the lowest of these: work-location Frechet elasticity for low-skill workers, $\eta_l = 2.07$ in his notation, to obtain $\zeta = 5.77$.[67]

[67] Specifically, let $\tilde{\zeta}$ given Frechet elasticity. Denote the expected utility of home-work pair ij for group g by $\overline{\Omega}_{ij}^g \equiv \theta_i^g + \overline{U}_{ij}^g + \omega_j^g$;

This nested logit demand system, as calibrated, now means commuters less willingly change home or work locations than they do modes. As a result, the planner has somewhat less scope for optimization of minibus passenger demand across routes, resulting in slightly lower welfare gains in the third line of Online Appendix Table E.4.

Agglomeration Spillovers

Fourth, I augment the model with agglomeration spillovers. In particular, I let wages in a location increase with total local employment $N_j \equiv \sum_{i,m,g} N^g \pi^g_{ijm}$ according to $w^g_j \equiv \xi^g_{0j} N_j^{\xi_1}$. The equilibrium conditions are then identical to those in the model in the main text, except with the endogenous wages w^g_j in place of the exogenous ω^g_j. Following Ahlfeldt et al. (2015), I set $\xi_1 = 0.07$ and now calibrate the ξ^g_{0j} to match locations' employment by workplace. For tractability, I assume that associations' bus supply choices do not account for the respective effects on wages w^g_j, but both the fare bargaining and the planner's solution do take agglomeration spillovers into account. In the fourth line of Online Appendix Table E.4, I find that accounting for agglomeration spillovers increases the gains from the planner's optimization of minibus demand and supply, as the planner concentrates workers in ex-ante more productive locations and thereby further pushes up wages.[68]

Non-Rush-Hour Demand

Fifth, I consider how the social planner might adjust demand and supply outside of rush hour. Approximately 84% of work commutes, as reported in the 2013 Cape Town Household Travel Survey, occur during the 6-10am rush hour, which I study and simulate in the main text. Outside of rush hour, non-work trips will make up a much more sizable share of total demand. I could capture non-work trips by incorporating a framework along the lines of Miyauchi et al. (2022); in this section, I provide an indication of how the welfare gains from optimization of minibus demand and supply might change during non-peak hours by solving the planning problem under commuter inflows N^g only three-fourths as high as in the baseline model. In the last row of Online Appendix Table E.4, I find even larger gains from optimization under this alternative, lower-demand scenario. Intuitively, due to the Market Mohring Effect, minibus routes suffer even longer wait times when demand is lower, which increases the gains from growing the "scale" of minibus commuting.

---

then, the Frechet elasticity satisfies $\tilde{\zeta} = \frac{\partial \log\left(\pi^g_{ij}/\pi^g_{kl}\right)}{\partial \log \overline{\Omega}^g_{ij}}$. In my logit demand system, instead,

$$\frac{1}{\zeta} = \frac{\partial \log\left(\pi^g_{ij}/\pi^g_{kl}\right)}{\partial \overline{\Omega}^g_{ij}} = \frac{\partial \log\left(\pi^g_{ij}/\pi^g_{kl}\right)}{\partial \log \overline{\Omega}^g_{ij}} \frac{1}{\overline{\Omega}^g_{ij}}$$

where the final equality uses the chain rule. Thus, a parameter ζ in my model equivalent to the Frechet parameter $\tilde{\zeta}$ must satisfy $\zeta = \frac{\overline{\Omega}^g_{ij}}{\tilde{\zeta}}$. I calculate the mean realized value of $\overline{\Omega}^g_{ij}$ in my baseline model, weighted by the commuter populations of each skill g commuting from i to j, and divide by $\tilde{\zeta} = \eta_l = 2.07$ from Tsivanidis (2023) to obtain $\zeta = 5.77$.

[68] To calculate equivalent variation, I find the proportionate change in a skill group's baseline wages w^g_j, at baseline values of $\{\boldsymbol{b}, \boldsymbol{\tau}, \boldsymbol{Q}, \boldsymbol{L}, \boldsymbol{T}, \boldsymbol{\kappa}, \overline{\eta}, \boldsymbol{H}, \boldsymbol{T_P}\}$, that leaves the average commuter equally well off as in the counterfactual.

Table E.4. Robustness: Social Planner Optimum

		Δ Mode Share					
		Minibus		Car		%Δ Welfare	
Case	*Skill:*	*Low*	*High*	*Low*	*High*	*Low*	*High*
Higher μ		0.04	0.03	-0.01	-0.01	0.16	0.08
Higher ρ		0.04	0.03	-0.01	-0.01	0.15	0.07
Nested Logit		0.04	0.03	-0.01	-0.01	0.14	0.07
Agglomeration		0.04	0.03	-0.01	-0.01	0.21	0.16
Non-Rush Hour		0.04	0.03	-0.01	-0.02	0.17	0.09

Notes: Table summarizes the social planner optimum under five alternative parameter values or modeling assumptions, including (i) a 25% higher queueing efficiency μ; (ii) a 25% higher minibus arrival efficiency ρ; (iii) nested logit commuter demand, with a nest for each home and work location pair comprising the three modes; (iv) agglomeration spillovers to wages, i.e. wages which increase with location employment, using the agglomeration elasticity in Ahlfeldt et al. (2015); and (v) "non-rush hour" commuter demand only 75% as high as in the baseline, to capture non-rush-hour conditions. For each case, I re-calibrate amenities and wages (or, in the case of agglomeration, the exogenous component ξ^g_{0j} of wages) as well as the association bargaining power β and driver wage $\overline{\omega}$, as in Section V. I then re-solve for a new baseline equilibrium as well as the new social planner optimum under each alternative parameter value or modeling assumption. Outcomes summarized in this table include the changes in the minibus and car mode shares by skill group as well as the percent change in group-level welfare, measured as equivalent variation, relative to the status quo.

E.4 Alternative Margins for Welfare Gains

In this section, I explore alternative margins through which the welfare gains from optimization of minibus demand and supply might be realized or increased. In particular, I consider alternative forms of bargaining in which the city government could engage as well as the scope for the planner to attain greater gains by having buses depart "early," before reaching capacity.

Nash Bargaining over Fares and Bus Supply

To explore whether a more complete form of private-public partnership between associations and the government might come even closer to achieving the optimum than the current arrangement of fare bargaining alone, I model a scenario in which associations and the government bargain over both fares and bus supply simultaneously. In particular, I assume that equilibrium bus supply b_{ij} and fares τ_{ij} on each route ij solve the Nash bargaining problem given by

$$\max_{\tau_{ij}, b_{ij}} \left\{ \beta \log \Pi_{ij} + (1-\beta) \overline{\Omega}_{ij} \right\}. \tag{E.4}$$

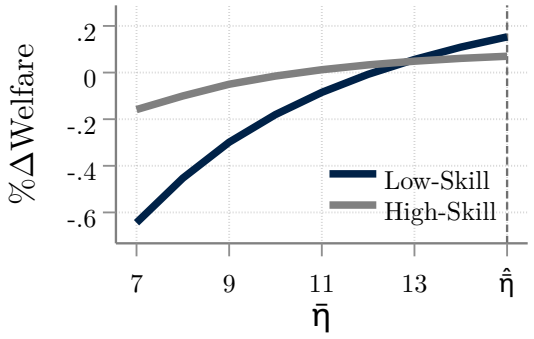
For the sake of brevity, I avoid restating the equilibrium conditions and formal equilibrium definition. Instead, I note that equilibrium now consists of a vector $\{\boldsymbol{b}, \boldsymbol{\tau}, \boldsymbol{\pi}\}$ that satisfies the two first-order conditions associated with (E.4) as well as the usual commuter demand system (8). Online Appendix Table E.5 compares the percentage changes in total minibus demand and supply as well as the welfare gains from the social optimum, relative to the decentralized equilibrium in my baseline model and in this alternative model with simultaneous

Table E.5. Social Optimum vs. Equilibrium with Bargaining Over Fares and Bus Supply

		%Δ in. . .			
				Welfare	
Model	*Skill:*	Minibus Passengers	Minibus Supply/Passenger	*Low*	*High*
Bargaining over Fares Only (Baseline)		12.6	81.6	0.15	0.07
Bargaining over Fares and Bus Supply		11.1	40.7	0.12	0.05

Notes: Table displays outcomes under the social planner optimum, relative to two different models of decentralized equilibrium market structure. The first line pertains to the baseline model, where the government and each association Nash-bargain over fares only, and associations then choose bus supply to maximize profit. The second line pertains to an alternative model where the government and each association Nash-bargain over both fares and bus supply. For each model, the table displays the percent change in total minibus passengers from the decentralized equilibrium to the social optimum, measured as the sum over routes of arrival rates λ_{ij}, the percent change in minibus supply per passenger, measured as the sum over routes of b_{ij} divided by the sum across routes of λ_{ij}, as well as the percent change in group-level welfare, measured as equivalent variation, in the social optimum, relative to the decentralized equilibrium. Note that the social optimum itself is independent of the assumptions on bargaining.

Figure E.2. Social Planner + Leaving Less-Than-Full: Welfare Gain



Notes: Figure displays the welfare gains which would result if the social planner could impose that buses leave less-than-full. In particular, the vertical axis displays the gains from the socially optimal commuter choices and bus supply presented in Section VII, with different values of minibus capacity $\overline{\eta}$ on the horizontal axis implemented by the planner along with the optimal choice probabilities and bus supply. Welfare gains are measured relative to the decentralized equilibrium with the status-quo bus capacity $\hat{\overline{\eta}} = 15$.

bargaining over fares and bus supply. In this alternative model, which could approximate a world where the government contracts on both price and quality, the status quo indeed comes even closer to the social optimum; the social planner does not need to increase bus supply as starkly anymore, and so the welfare gains from the optimum are even smaller.

Less-Than-Full Bus Departures

I now seek to understand whether the planner could attain additional gains by having buses depart before reaching a full load of passengers, which would, of course, further reduce loading times. As a suggestive exercise, I recalculate the social planner optimum under lower bus capacities $\overline{\eta}$; because I do not adjust

operations costs χ_{ij}, this exercise simulates the gains from having buses depart, on average in my probabilistic departure framework, with fewer passengers onboard. In Online Appendix Figure E.2, I then display the welfare gains from these social optima with "early" departures, relative to the decentralized equilibrium with the status-quo bus capacity $\hat{\bar{\eta}} = 15$. Additionally imposing lower bus capacities, i.e. earlier departures, turns out to always decrease the gains from the planner's optimization of demand and supply. Intuitively, the planner already scales up demand to leverage the Market Mohring Effect and reduce loading times to manageable levels, such that the additional gains from requiring buses to depart even earlier do not counterbalance the wasted operations costs from departing with partially-empty buses.