

Privatized Provision of Public Transit*

Lucas Conwell [†]

First Version: October 2022

This Version: July 2024

Abstract

Workers in developing countries waste significant time commuting, and gaps in public transit constrain access to productive jobs. In many cities, privately-operated minibuses provide 50–100% of urban transit, at the cost of long wait times and poor personal safety for riders. Can policymakers improve upon the privatized provision of public transit via subsidies, which leverage increasing returns in wait times, or technological upgrades? I build a micro-founded model of privatized shared transit where minibuses load passengers from a queue. I then estimate the model with newly-collected data on minibus and passenger queues in Cape Town and stated user preferences for exogenously-varied commute attributes. I find that governments should subsidize minibuses and their passengers because neither internalize their beneficial spillovers to wait times. This optimization shortens queues, fills buses faster, and particularly benefits low-skill workers. Government actions to expand minibus stations or improve security bring even more substantial welfare gains.

*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 Guadalupe Bedoya, Yizhen Gu, Andrii Parkhomenko, Ian Savage, and Román David Zárate for their insightful discussions. I am grateful to Treb Allen, Richard Blundell, Richard Carson, Aureo de Paula, Mateus Dias, Fabian Eckert, Simon Fuchs, Gabriel Kreindler, and Edouard Schaal for their suggestions 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 additionally thank my former Yale colleagues for their extensive input, including Lorenzo Caliendo, Will Damron, Cecilia Fieler, Claudia Gentile, Antonia Paredes-Haz, Sam Kortum, Ryungha Oh, Mark Rosenzweig, Nicholas Ryan, Matthew Schwartzman, Sam Slocum, Michael Sullivan, Kaushik Vasudevan, and Trevor Williams. I thank seminar participants and audiences at the University of Cologne, Bocconi University, CREI, Vrije Universiteit Amsterdam, University of Bristol, Hertie School, University College London, Georg-August-Universität Göttingen, the NBER Economics of Infrastructure Investment: Public Transportation Conference, the UEA European and North American Meetings, the ADB Conference on Infrastructure and Urban Development in the Developing World, EEA Annual Congress, Católica Lisbon School of Business & Economics, Toulouse School of Economics, TPUG Session at ASSA, CRED Workshop on Regional and Urban Economics, World Bank Urbanization and Poverty Reduction Research Conference, BREAD @ UCSD, the German Development Economics Conference, and the Berlin Quantitative Spatial Economics Seminar. I thank Philip Krause, Aslove Mateyisi, and Mokgadi Mehlaope from GoAscendal/GoMetro for their herculean efforts to obtain permission for and organize the data collection. Finally, I acknowledge generous financial support from the Yale Economic Growth Center and 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

Workers in lower-income countries waste significant time commuting (Kreindler (2022) and Akbar et al. (2023b)), and these often-unsafe journeys constrain their access to the most productive jobs (Tsivanidis (2023) and Zarate (2024)). Simultaneously, due to limited fiscal capacity and increasing sprawl, governments often struggle to expand formal public transit at pace with urbanization (Balboni et al. 2020). Into these gaps have emerged vast networks of privately-operated minibuses, which, today, provide 50 to 100% of urban transit in many developing-world cities (Tun and Hidalgo (2022)). In my context of Cape Town, minibus passengers spend one-third of their commutes in queues to board and then wait for the bus to fill up and depart. Commuters benefit from atypically new, roadworthy minibuses but suffer a high crime risk. These chaotic networks offer a lifeline of connectivity to the urban poor, yet economists have rarely studied this form of *privatized shared transit*.

In this paper, I fill this gap and ask how policymakers can improve upon the privatized provision of public transit. Could appropriate subsidies leverage increasing returns in the physical minibus loading process to optimize Cape Town’s minibus supply and passenger waits? Alternatively, could deregulated pricing or free minibus entry approximate this social optimum? Finally, I consider whether technological upgrades, such as expanded stations or security guards, can similarly alleviate spatial mismatch between workers and jobs and reduce emissions.

To do so, I introduce a model of privatized shared transit. Passenger and bus queues generate increasing returns in the wait times which so often plague privatized shared transit (Cervero and Golub (2007)) and transport markets more generally (Brancaccio et al. (2020b) and Fuchs and Wong (2024)). I then nest this supply framework within a canonical quantitative urban model of commuting (Ahlfeldt et al. (2015)). I collected new data on Cape Town’s minibus loading process, itself emblematic of global industry practices (Kerzhner (2022)), with which I quantify queueing efficiency. Simultaneously, I extend advances in stated preference surveys to identify user preferences for difficult-to-measure yet easily understandable commute attributes, e.g. security guards. The estimated model reveals that optimal subsidies decrease passenger waits in the queue, fill buses faster, and particularly benefit low-skill, lower-income commuters. However, additional queueing areas and government-provided station security guards generate even larger welfare and emissions gains.

My model integrates a privatized shared transit sector into a quantitative urban model of demand, the two of which interact through bus and passenger queues. A minibus association on each distinct origin-destination route chooses the supply of buses, which arrive to the station to load passengers from a queue. Commuters with heterogeneous incomes choose a mode of transport as well as home

and work locations based on factors such as commute times and safety. At the heart of the model, a two-sided queue determines the wait times of buses and passengers. Passengers first queue to board buses and subsequently wait on these buses, which depart only when full. Passenger *queueing* times fall with bus supply, and bus *loading* times fall with demand, even in an extension where I permit buses to depart less-than-full. In the spirit of the so-called Mohring (1972) Effect, whereby wait times for *government*-provided transit fall with scale, I term the two aforementioned spillovers the *Indirect* and *Direct Market Mohring Effects*. Thus, the physical constraints inherent to the minibus loading process introduce a form of increasing returns which equilibrium fares generically fail to correct.

I collected two forms of primary data in Cape Town to directly measure queues and quantify consumer preferences for typically unobserved commute attributes. First, enumerators observed passenger and bus queues on a random sample of 44 minibus routes, from which I measure passenger arrivals and wait times. Second, I employ stated preference methods to generate exogenous variation in commute choices, a particularly tangible and familiar context for respondents. In my survey, 526 respondents chose hypothetical minibus commute options with exogenously-varied travel times, costs, and quality improvements. Preferences across modes, not only minibuses but also car and “formal” public transit, come from a separate city-run stated preference survey.

With the help of these two datasets, I estimate the two-sided queueing model and the commuter demand system. I directly estimate the efficiency of the passenger queueing process from the relationship between bus dwell times per arriving passenger and queue length, the latter of which can exceed 60 passengers. Unobserved interruptions to this loading process, e.g. due to weather or special events, could lengthen dwell times and thus threaten identification. I instrument for year-2022 dwell time per passenger with year-2013 commute start times, likely uncorrelated with interruptions nine years later. From commuters’ stated preferences, I estimate a discrete choice model that yields commuters’ mode-specific utility costs as well as values of time and security in dollar equivalents. The high-skill, in particular, perceive such a high crime risk that they would pay a full \$2.75 per commute for station security guards.

Finally, I employ the estimated model to demonstrate that governments can optimize the privatized provision of urban transit through supply and demand subsidies. My model features two key inefficiencies, or sources of increasing returns. First, congestion in bus arrivals generates a spillover from bus supply to passenger queueing times. Second, when multi-passenger buses load via a queue, their loading times decrease with the scale of demand. I show that route-specific subsidies to minibus associations and commuters correct these spillovers and increase welfare by 1.35 percent for low-skill workers and by 0.4 percent for the high-skill. The planner would raise demand on low-wage, suburban routes where buses fill slowly and simultaneously takes advantage of mild bus

congestion to increase supply on busy, high-wage routes. Decreased commute times and emissions more than outweigh a mild drop in average wages earned by commuters.

Among alternative policies, only technological upgrades, loosely defined, outperform the optimal subsidies in welfare terms. In particular, monopoly pricing by associations, as well as free minibus entry, each uniformly adjust supply and demand across routes and thus fail to raise welfare. In contrast, a second queue for each route and government-provided station security guards to ward off robberies and assaults each increase welfare by up to 2.4 percent. Thus, improved privatized transit could help solve the transport problems of a host of rapidly growing, resource-poor African cities similar to Cape Town.

Related Literature

A long tradition in economics highlights the advantages of bus transit: low fixed costs (Meyer and Wohl (1965)), flexibility (Glaeser (2020)), and, indeed, the range of vehicle sizes which can profitably serve lower-demand markets (Walters (1979), Mohring (1983), White (1987), and Oldfield (1988)). A more recent literature employs what are known as quantitative spatial models to study the effects of such transport connections on the wider economy (Allen and Arkolakis (2014), Ahlfeldt et al. (2015), Donaldson and Hornbeck (2016), Donaldson (2018), Monte et al. (2018), Hebligh et al. (2020), and Nagy (2023)). The most closely-related papers quantify how travel time reductions from bus rapid transit or subway lines translate into spatial sorting and job matching (Tsivanidis (2023) and Severen (2023)), informality (Zarate (2024)), or gentrification (Balboni et al. (2020) and Warnes (2021)). I contribute the first model of the privatized transit sector and then study policies whose impact depends crucially on the associated changes in bus supply and wait times.

Another strand of the literature explicitly considers feedback between transport costs and agents' decisions via road congestion (Duranton and Turner (2011), Kreindler (2022), Barwick et al. (2022), and Akbar et al. (2023a)), network effects, and environmental externalities (Almagro et al. (2024)). Recent advances then characterize optimal road (Fajgelbaum and Schaal (2020) and Allen and Arkolakis (2022)), public transit (Kreindler et al. (2023)) and ride-sharing (Almagro et al. (2024)), commuting (Fajgelbaum et al. (2021)), ocean shipping (Brancaccio et al. (2023)), or multi-modal networks (Fuchs and Wong (2024)). My supply-side model instead highlights externalities key to privatized transit and then characterizes the social optimum for developing countries' most common commute mode.

Methodologically, I build on models of transport wait times as well as a nascent literature on stated preference estimation. The former typically employ matching (Brancaccio et al. (2020a)) to study markets such as ride-hailing (Castillo (2022)) or ocean shipping (Brancaccio et al. (2020b)).

Brancaccio et al. (2024), who study ships' wait and service times in port, and my paper both introduce queueing models which provide one micro-foundation for matching. On the empirical side, stated preference surveys have increasingly provided exogenous variation to estimate structural models of long-term care (Ameriks et al. (2020)) or marriage decisions (Andrew and Adams-Prassl (2023)). I extend the stated preference methodology to a particularly concrete, familiar context where respondents can likely accurately predict their behavior.

I structure the remainder of the paper as follows. In Section II, I describe my data, both newly collected and from existing sources, and the context. In Section III, I discuss a series of facts to rationalize my modeling choices and counterfactuals. 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 discuss the welfare gains from alternative transport policy interventions 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 employed in the paper.

Newly-Collected and Household Survey Data

I filled the gap in existing micro-data on privatized transit with a custom two-part data collection effort in Cape Town. First, I observe 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. 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. With these observations in hand, I imputed passengers' queueing times, which I set aside as an over-identification check. The counts covered a two-stage cluster sample of $N = 44$ minibus routes in Cape Town, where I sampled origin stations and then routes that originate from sampled stations. I pair these observations of the loading process with fares recorded by enumerators on board minibuses.

Second, I quantified commuter demand via a stated preference survey designed to estimate the monetary values of time and quality as well as commuters' price sensitivity. 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 1. Options varied in five attributes: cost, travel time, and three binary quality improvements, which I chose based on commuter concerns in

FIGURE 1. STATED PREFERENCE SURVEY CHOICE SET BETWEEN MINIBUS OPTIONS

Q1.1	Option 1.1.1	Option 1.1.2
Cost	R18.00	R6.00
Travel Time	50 Minutes	50 Minutes
Security	Security at taxi rank 	No security at taxi rank 
Driver Behaviour	Adheres to speed limit 	Exceeds speed limit 
Bus Loading	Enough seats for all passengers 	Overloaded: more passengers than seats 

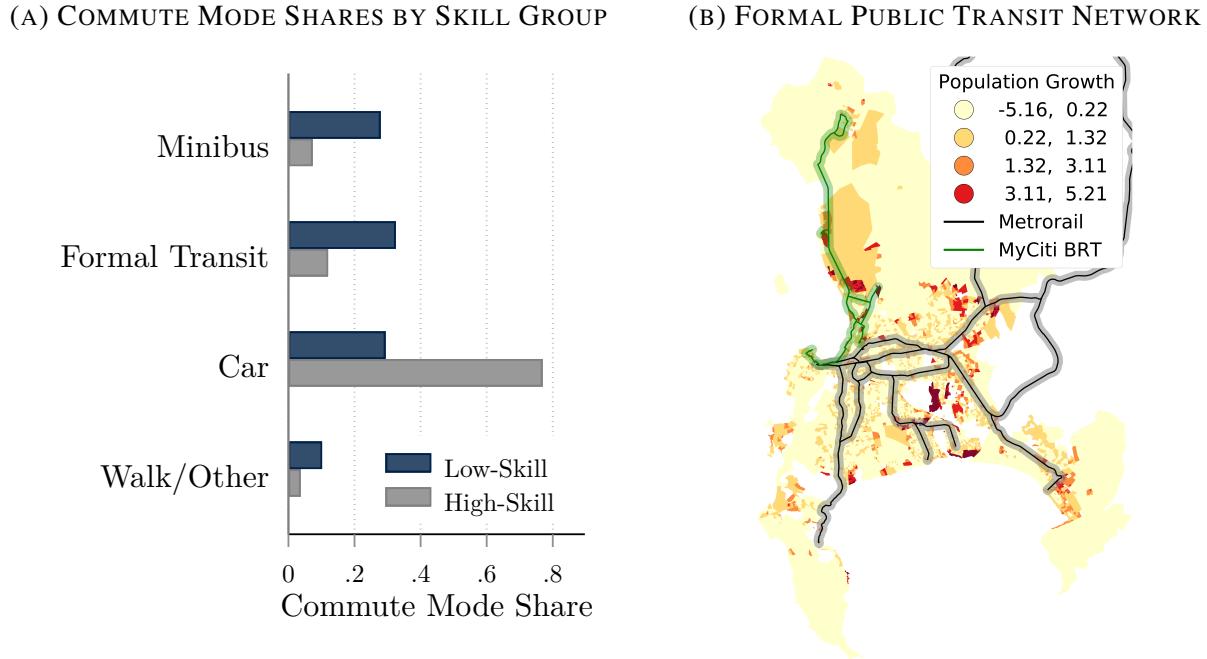
Notes: This figure shows an example of a choice set from my stated preference survey, consisting of two hypothetical minibus commute alternatives, from which respondents indicated their preferred option. The rows list the attributes associated with each option, which vary exogenously across choice sets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

past surveys. The latter included security guards at the publicly-owned, shared minibus stations, driver adherence to speed limits, and whether the minibus loads more passengers than seats. I chose the levels of the attributes with a *d-efficiency* algorithm that maximizes parameter precision (Rose and Bliemer (2009)), and respondents completed one of two randomized “blocks” of five choice sets. After a pilot survey, I reduced the number of choice sets and alternatives per set to maintain respondent attention.

To conduct interviews, enumerators randomly approached respondents at one intermodal transport hub and two minibus stations on weekdays. This frame, chosen to economize on resources, produced a sample representative of the entire Cape Town commuter population in terms of age, gender, education, and income. However, my survey over-sampled minibus commuters, as detailed in Online Appendix B.2.6. Therefore, I employ this data only to estimate relative preferences for different *minibus* attributes and later attempt to quantify disparities in preferences between my sample and the population. Additionally, I collected respondents’ *revealed* preferences, i.e. actual commute modes, for later comparison to their stated preferences.

Finally, I use representative household survey data collected by the City of Cape Town. Most importantly, a city-run stated preference survey varies the mode of transport, namely car, formal public transit, or minibus, rather than commute quality, as in my survey. Additionally, for a larger sample of residents, the same data provide home and workplace locations as well as incomes.

FIGURE 2. DIFFERENT 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 formal transit modes with dedicated infrastructure in Cape Town, namely MyCiti bus rapid transit and Metrorail commuter trains. Shading indicates population growth from 1996–2011 at the small area layer level.

Minibuses in Cape Town

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 South Africa. Figure 2a displays the shares of commuters in Cape Town on each mode of transport. Throughout the paper, I distinguish between low-skilled workers, or those with less than a university education, and the high-skilled. A full 28% of low-skill workers and 7% of high-skill workers commute via minibus; the overwhelming majority of high-skill commuters instead drive to work. Around one-third of low-skill commuters use limited publicly-provided “formal” transit alternatives. These include infrequent Golden Arrow buses running in mixed traffic, higher-speed MyCiti bus rapid transit, and Metrorail commuter rail lines. However, the latter two networks, overlaid on recent population growth in Figure 2b, miss many fast-growing suburbs.

The minibus market consists of a large number of private firms who own, on average, less than two buses each (Woolf and Joubert (2013)). Firms pay an entry fee to an owner *association* to operate on a specific origin-destination pair, or *route*. Associations, in turn, obtain exclusive rights to regulate firm entry and fares on one or more routes from city government (City of Cape Town (2014) and

Kerr (2018)). Entry fees furnish the lion’s share of associations’ revenue; as a result, attempts to infringe on these state-sanctioned monopolies occasionally provoke violent retaliation (Schalekamp (2017) and Kerr (2018)). Associations’ chosen fares, in turn, require city government approval of the “portion of monthly [passenger] income spent on public transport” (City of Cape Town (2014) p.65).¹ Contrary to popular belief, a similar mix of sophisticated collective organization and government regulation defines minibus markets across the African continent (Kerzhner (2022, 2023)).

III. FACTS ABOUT THE MINIBUS MARKET

I now present six facts about the minibus sector in Cape Town. As in cities from Djibouti to Nairobi to Lilongwe, buses depart at random, unscheduled times from so-called *taxis ranks* (Kerzhner (2022)). Rush-hour passengers also typically board at these large minibus stations, which feature clearly marked loading lanes for each route. Henceforth, I deviate from local terminology and call these facilities *minibus stations*.

Fact 1. *Passengers wait in sometimes-sizeable queues to board minibuses at the station where a route originates.*

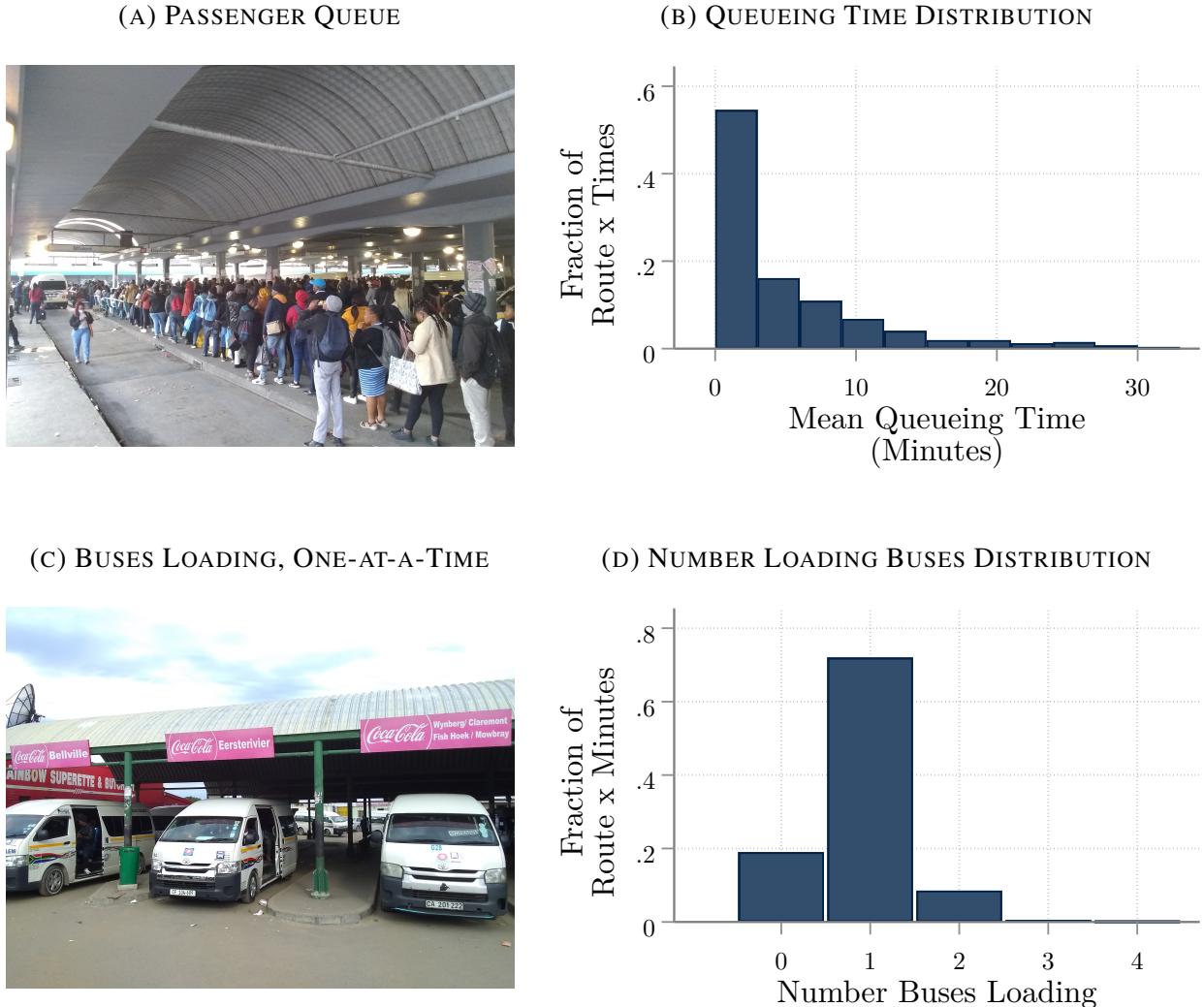
Long rush-hour queues, as pictured in Figure 3a, loom large in the public imagination. Local journalists, for example, complain that “queues, especially during certain times of the day, are impossibl[y]” long (Theway (2018))—but these waits turn out not to afflict all routes equally. The histogram in Figure 3b displays passengers’ average wait times in the queue, imputed across routes and five-minute time blocks in my station count data. Queues only arise on around half of routes and time periods. Nonetheless, the distribution displays a long right tail of routes where this *queueing time* almost exceeds the average *total* minibus commute time of 36 minutes. To account for the substantial time thus lost by commuters, my model features a first-in-first-out passenger queue for each route.

Fact 2. *Exactly one minibus per route loads passengers during 70% of peak commute times; during most of the remaining time, the loading bays remain empty.*

Figure 3c displays the typical case where one minibus at a time loads in each route’s loading bay. The histogram in Figure 3d of the number of buses on the *same* route which load simultaneously, across routes and minutes, confirms the ubiquity of this pattern. Even during the peak hours I study, the loading areas nonetheless remain empty almost one-fifth of the time. My model thus augments

¹Virtually all firms join associations (Antrobus and Kerr (2019)). Additionally, the law requires individual firms to obtain an effectively pro-forma government permit and operate a vehicle with one of several approved seat capacities (Jobanputra (2018) p. 290). However, up to half of firms lack these permits (City of Cape Town (2014) p. 77).

FIGURE 3. MINIBUS LOADING PROCESS AT ORIGIN STATION



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 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 (C) displays the Mfuleni minibus station with one bus per route loading in the corresponding designated loading lane. Panel (D) displays the distribution of the number of minibuses loading at the origin station on a specific minibus route, over minibus routes and minutes. Images by author, June 2022.

the aforementioned passenger queue with a bus queue: one bus loads at a time, and, thereafter, the next arrives only with some delay.

Fact 3. *15-passenger vans account for 94% of licensed minibuses, and 96% of minibuses depart with a full load of at least 15 passengers.*

Though the law allows for several discrete bus sizes, the 15-passenger-plus-driver variant dominates

in practice (Jobanputra (2018) p. 290). Furthermore, minibuses in Cape Town rarely depart with empty seats and thus emulate the “possibly inefficient” (Kerr (2018)) *fill-and-go* practice near-universal among privatized shared transit providers worldwide (Cervero and Golub (2007)). In addition to the aforementioned *queueing time*, this *loading time* adds a second component to passengers’ total wait time for departure. In consequence, I impose a single exogenous bus size in my model and require buses to reach this exogenous capacity, in expectation, before they depart.²

Queueing and loading times, far from inevitable, exhibit a particular form of increasing returns to scale:

Fact 4. *Passengers wait less on busier minibus routes, both because queueing times fall with bus supply and because loading times fall with demand.*

Figure 4a plots the relationship between, on the horizontal axis, the rate of newly-arriving buses for a given route and hour and, on the vertical axis, mean queueing times. The latter fall with bus supply thanks to shorter empty periods between buses. Furthermore, Figure 4b demonstrates that faster passenger arrivals, on the horizontal axis, fill minibuses more quickly and thus alleviate the long loading times on the vertical axis.

These increasing returns mirror the famous Mohring (1972) Effect: higher demand for government-provided, formal public transit justifies higher frequencies and thus lower wait times. The aforementioned beneficial effect of bus supply on queueing times drives what I term, in the case of privatized shared transit, the *Indirect Market Mohring Effect*. The latter negative link between demand and loading times, in turn, underlies a *Direct Market Mohring Effect*. The latter elasticity alone, at -0.74 , slightly exceeds estimates of the elasticity of formal transit wait times with respect to demand in Parry and Small (2009).³ However, the Market Mohring Effects arise entirely through the decentralized choices of associations and passengers. In my policy counterfactuals, I thus consider how optimal government subsidies or changes in market structure might leverage the Market Mohring Effects to reduce wait times.

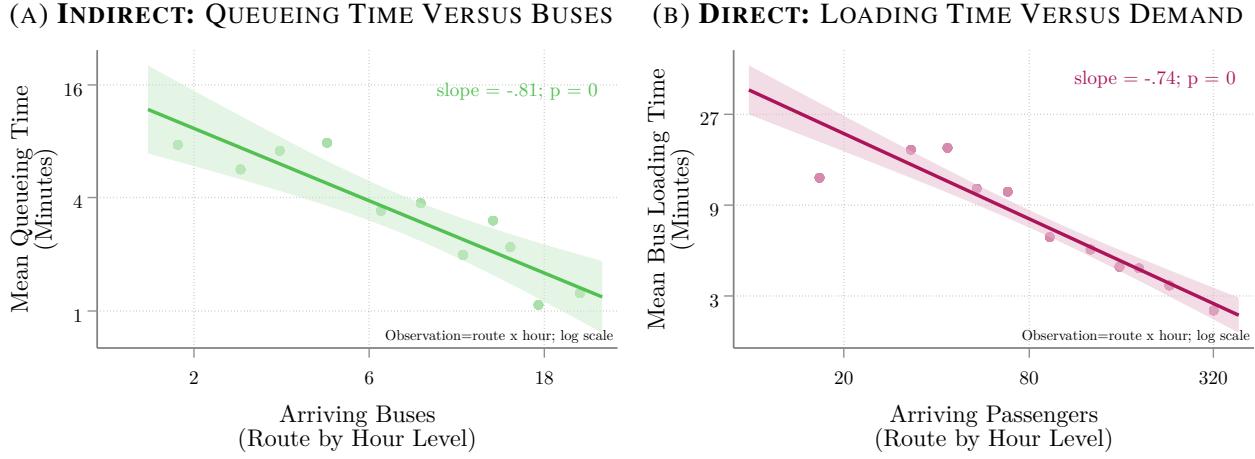
However, commute utility depends on a range of factors beyond mere time:

Fact 5. *Commuters in a past satisfaction survey rated security from crime second-most negatively among all minibus attributes.*

²In Online Appendix E.3, I endogenize the number of passengers with which buses depart; my model replicates the status quo, in that buses on nearly all routes continue to leave full. The passenger experience, once underway, more closely resembles that of scheduled transit. Most routes follow a “line-haul” mode of operation during peak hours: they travel on highways to their destination and do not expect to pick up additional passengers.

³The 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 4. MARKET MOHRING EFFECTS



Notes: Panel (A) displays binned scatterplot and best fit line for the log-scale relationship between the number of arriving buses per hour for a given minibus route and hour and mean passenger queueing time imputed from the station count data. Panel (B) instead displays the relationship between arriving passengers and mean minibus loading times at the route by hour level, also from the station counts.

Indeed, in a 2013 survey, security led the list of minibus-related grievances that included road safety, crowdedness, cleanliness, timetable adherence, ease of use, and distance to the stop, as displayed in Online Appendix Figure A.1. The fundamental public-good nature of security in shared public spaces such as minibus stations in and of itself rationalizes state intervention. Only a select few stations currently post guards to alleviate the risk of theft or assault, so I consider a counterfactual of government-financed security guards across all minibus stations.

IV. A THEORY OF PRIVATIZED SHARED TRANSIT

In this section, I build a model of the privatized shared transit sector. On the supply side of the model, minibus associations set bus entry and fares, after which minibuses load passengers in a two-sided queue. Demand follows a full quantitative spatial framework, whereby commuters with heterogeneous incomes optimally choose their home location, work location, and mode of transport. I lay out the environment, discuss the queueing technology as well as the problems of each type of agent, define equilibrium, and then derive its efficiency properties.

Environment

I consider a city made up of a finite number of locations indexed by $i, j \in \{1, \dots, I\} \equiv \mathcal{L}$ and three types of agents: minibuses, minibus associations, and commuters. Time in my dynamic model is

continuous, and commuters discount the future at rate r . Each origin-destination pair, or what I term a *route*, constitutes a separate minibus market in which a route-specific association acts as a monopolist. Minibuses on a route wait in reserve until they arrive to the passenger queue and load commuters for a trip. The single minibus association for each route freely chooses the supply of minibuses and, endowed with bargaining power β , Nash-bargains fares with city government. Fixed masses $\{N^g\}$ of commuters of skill $g \in \{\text{low } (l), \text{high } (h)\}$ 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 one mode $m \in \{\text{minibus } (M), \text{formal transit } (F), \text{car } (A)\}$ for a single commute to work. They collect a fixed wage ω_j^g upon arrival. I leave the parameters related to formal transit and cars as exogenous and solve the model exclusively in steady-state.

A Two-Sided Minibus Queue

I model the frictional process by which minibuses load passengers on a given route as a form of two-sided queue. Buses wait in reserve to arrive to a route's origin minibus station, where passengers wait for the bus's arrival in a route-specific queue.

First, I tailor the one-sided M/M/1 queueing model with service interruptions in Avi-Itzhak and Naor (1963) to model the passenger side of the queue. Passengers on origin-destination route ij arrive to the route's first-in-first-out queue at Poisson rate λ_{ij} , itself a function of commuter demand. Whenever a bus loads in the queueing area, passengers from the front of the queue board the bus at Poisson rate μ . This *queueing efficiency*, inversely proportional to the average time passengers take to board, reflects station infrastructure but also behavioral and cultural factors.

Second, an approximation of a bus queue feeds minibuses to this standard one-sided queue of passengers. Suppose that the route-specific association supplies b_{ij} minibuses. Then, after a bus departs and leaves the queueing area empty, another arrives at a rate $\lambda_{ij}^B \equiv g(b_{ij}) \varsigma_{ij}$. A larger bus supply “reserve” ensures that a new bus can quickly take a departed bus's place, i.e., $g'(\cdot) > 0$. Nonetheless, $g(\cdot) < \infty$ because vehicular congestion en route to and within the minibus station precludes instantaneous arrivals; ς_{ij} accounts for unobserved variation in arrival efficiency. Furthermore, to idiosyncratic disruptions to the loading process, e.g. weather or special events, the newly-arrived bus can load passengers only upon receipt of an unobserved “loading shock” at Poisson rate ι_{ij} . The bus departs at rate $\tilde{\lambda}_{ij}/\bar{\eta}$, where $\tilde{\lambda}_{ij}$ denotes the expected outflow of passengers from the queue onto the bus and $\bar{\eta}$, bus capacity. Probabilistic departures ensure that the average bus departs with a full load of $\bar{\eta}$ passengers, as laid out in Fact 3, but simultaneously preserve the stationarity of the model.⁴

⁴In Online Appendix E.3, I allow associations to choose the equivalent of $\bar{\eta}$, and the overwhelming majority of routes continue to depart with full loads, in the status quo and in my primary counterfactual.

Passenger demand, λ_{ij} , and the bus supply b_{ij} then jointly determine the total time passengers wait in the queue and for the bus to fill. In Appendix A.1, I leverage an isomorphism with the queueing model of Avi-Itzhak and Naor (1963) to derive intuitive expressions for these waits. In particular, the probability P_{ij}^L that a bus loads passengers at any given moment on a route increases with bus supply,

$$P_{ij}^L = 1 - \frac{\lambda_{ij}}{\bar{\eta}} \left(\frac{1}{g(b_{ij})} + \frac{1}{\iota_{ij}} \right). \quad (1)$$

Total wait time, in turn, consists of two components. First, passengers wait an expected queueing time,

$$Q_{ij} = \frac{\mu^{-1} + \frac{\bar{\eta} P_{ij}^L (1 - P_{ij}^L)^2}{\lambda_{ij}}}{P_{ij}^L - \frac{\lambda_{ij}}{\mu}} \approx \frac{1}{\mu P_{ij}^L - \lambda_{ij}}, \quad (2)$$

to board. Queueing time logically rises with demand, λ_{ij} , but falls, the more efficiently passengers board buses, μ , and the more often a bus loads in the station, P_{ij}^L . The latter channel underlies the Indirect Market Mohring Effect in Figure 4a. Second, because minibuses only depart once full, passengers also wait out the expected loading time,

$$L_{ij} = \frac{\bar{\eta} P_{ij}^L}{\lambda_{ij}}, \quad (3)$$

until departure.⁵ This latter wait time distinguishes minibuses from one-to-one technologies such as taxis and, crucially, falls with demand to generate the Direct Market Mohring Effect in Figure 4b. Finally, I note that the loading shocks and idiosyncratic variation in bus arrivals serve primarily to motivate my estimation strategy. Thus, in the remainder of this section and throughout the characterization of equilibrium and counterfactuals in Sections VI-VII, I impose

Assumption 1. *On all routes, buses load instantly upon arrival, $\iota_{ij} \rightarrow \infty$, and bus arrivals display no unobserved heterogeneity, $\varsigma_{ij} = 1$.*

Associations

Associations, which each possess exclusive rights to a single minibus route, make two key decisions. First, the association Nash bargains fares with city government, and second, conditional on fares, the association chooses bus supply to maximize its flow of profits. I discuss and solve the association

⁵The assumed bus departure rate, $\tilde{\lambda}_{ij}/\bar{\eta}$, combined with steady-state flow balance $\lambda_{ij} = P_{ij}^L \tilde{\lambda}_{ij}$, imply that buses depart at rate $\lambda_{ij} / (\bar{\eta} P_{ij}^L)$. By the properties of Poisson processes, the expected loading time equals the inverse of the rate of bus departure.

problem via backward induction.

Bus Supply

A minibus association's bus supply continuously runs trips on its designated route. Individual minibuses wait in a reserve pool of chosen mass b_{ij} , which a randomly chosen minibus leaves each time the route's loading bay empties. The bus then arrives to the origin minibus station, immediately—under Assumption 1—loads passengers for, on average, L_{ij} minutes, drives to the route's destination in a fixed travel time T_{ij} , and re-enters the reserve.

Associations earn revenue from fares, denoted by τ_{ij} , and pay operations costs and driver wages. Each trip on a route costs χ_{ij} in operations costs; drivers earn a fixed wage $\bar{\omega}$ per unit time. The flow of minibus association profits, Π_{ij} , equals the arrival rate of busloads times, in brackets, profits per trip, net of wages paid to drivers who wait in reserve:

$$\Pi_{ij} \equiv \underbrace{\frac{\lambda_{ij}}{\bar{\eta}}}_{\text{Busloads}} \left[\underbrace{\bar{\eta} \tau_{ij}}_{\text{Revenue}} - \underbrace{\chi_{ij}}_{\text{Ops. costs}} - \underbrace{\bar{\omega} (L_{ij} + T_{ij})}_{\text{Trip time}} \right] - \underbrace{\bar{\omega} b_{ij}}_{\text{Reserve costs}}. \quad (4)$$

Finally, associations then choose bus supply b_{ij} to maximize profits conditional on bargained fares, which yields the first-order condition

$$0 = \frac{\Pi_{ij} + \bar{\omega} b_{ij}}{\lambda_{ij}} \frac{\partial \lambda_{ij}}{\partial b_{ij}} - \frac{\lambda_{ij}}{\bar{\eta}} \bar{\omega} \left(\frac{\partial L_{ij}}{\partial \lambda_{ij}} \frac{\partial \lambda_{ij}}{\partial b_{ij}} + \frac{\partial L_{ij}}{\partial b_{ij}} \right) - \bar{\omega}. \quad (5)$$

The first term captures a beneficial equilibrium effect of bus supply increases, namely lower passenger wait times and thus higher demand, weighted in (5) by profits per passenger. The second term subtracts off the extent to which the accompanying higher loading times raise labor costs, and the third, wages paid to additional drivers as they wait to arrive to the station.

Fares

Before but in anticipation of this optimal supply choice, the association on each route separately Nash-bargains fares with city government, conditional on all other such pairwise bargains. This “Nash-in-Nash” solution concept, commonly employed to solve models of bilateral oligopoly with spillovers (Collard-Wexler et al. (2019) and Yürükoglu (2022)), here replicates Cape Town's government-mediated fare scheme. Associations, armed with bargaining power β , seek to maximize profits, Π_{ij} . In line with the official route approval guidance discussed in Section II, the city govern-

ment's objective equals average destination workplace income ω_j , net of the minibus fare.⁶ I assume that both sides receive zero payoff if bargaining fails, so fares, τ_{ij} , maximize $\Pi_{ij}^\beta (\omega_j - \tau_{ij})^{1-\beta}$. The resulting fares,

$$\tau_{ij} = \omega_j - \frac{1-\beta}{\beta} \frac{\partial \Pi_{ij}}{\partial \tau_{ij}}, \quad (6)$$

naturally rise with association bargaining power, β , but, more subtly, fall with ex-ante profitability. On these often shorter routes, associations forgo larger gains in the event of failed negotiations and thus accept lower fares.

Commuters

Commuters of skill g choose the combination of home location i , work location j , and mode m , either minibus, formal transit, or car, which offers the highest utility, $\theta_i^g + U_{ijm}^g + \omega_j^g + v\varepsilon_{ijm}$. Total utility comprises (i) the home location's exogenous amenity value θ_i^g (ii) the deterministic commute value U_{ijm}^g ; (iii) the work location's fixed wage ω_j^g ; and (iv) a Gumbel-distributed idiosyncratic preference ε_{ijm} with variance scaled by the parameter v .⁷

The commute value U_{ijm}^g depends on mode-specific utility, time, and monetary costs and linearly approximates a micro-founded commute model, detailed in Online Appendix C.3. In particular, minibus commuters enjoy commute value

$$U_{ijM}^g \equiv -\kappa_M^g - r\omega_j^g \left(Q_{ij} + L_{ij} + T_{ij} \right) - \tau_{ij}. \quad (7)$$

\uparrow \uparrow \uparrow \uparrow \uparrow
utility cost queueing time loading time travel time fare

The skill-specific utility cost κ_m^g reflects commuters' non-pecuniary taste for a particular mode as well as quality improvements, such as improved minibus station security. In turn, the *product* of the time preference rate r and the wage ω_j^g determines commuters' value of reductions in the total commute time in parentheses. Finally, commuters' sensitivity to the fare, τ_{ij} , charged varies inversely with the Gumbel scale parameter v . Commuters on other modes make similar tradeoffs. Formal transit commuters pay a utility cost κ_F^g as well as exogenous fares τ_{ijF} , wait a fixed headway H_{ij} , and travel a fixed T_{ijF} minutes to their destination, so that $U_{ijF}^g \equiv -\kappa_F^g - r\omega_j^g (H_{ij} + T_{ijF}) - \tau_{ijF}$. Car commuters similarly pay a utility cost κ_A^g as well as a fixed monetary cost τ_A and spend the same

⁶Note that, for simplicity, this average workplace income uses weights equal to a skill group's share of the aggregate city population, so that $\omega_j \equiv \left(\sum_{g'} N^{g'} \right)^{-1} \sum_g N^g \omega_j^g$.

⁷I do not allow workers to combine modes; in my stated preference survey, only 4.6% of minibus riders report also using formal buses or trains over the course of their commutes.

on-road travel time T_{ij} as minibuses; they thus enjoy commute value $U_{ijA}^g \equiv -\kappa_A^g - r\omega_j^g T_{ij} - \tau_A$.

Aggregate commuter demand then adheres to three choice-probability equations of the familiar Gumbel form. Skill- g commuters choose to commute by mode m from home i to work j with a probability π_{ijm}^g that satisfies

$$\pi_{ijm}^g = \exp\left(\frac{\bar{W}^g}{v}\right)^{-1} \exp\left(\theta_i^g + U_{ijm}^g + \omega_j^g\right)^{1/v}, \quad (8)$$

where \bar{W}^g denotes commuters' ex-ante expected Gumbel utility.⁸ Minibus demand translates into passenger arrivals,

$$\lambda_{ij} \equiv \sum_g N^g \pi_{ijM}^g, \quad (9)$$

so the scale of demand on a route follows directly from the exogenous fundamentals—amenities and wages—of the origin and destination.

Equilibrium and Efficiency

Equilibrium in my model depends on one set of universal parameters and another model geography of parameters specific to origins, destinations, or both. These parameters determine minibus supply, minibus fares, commuter choice probabilities, passenger arrivals, loading probabilities, and wait times, as follows.

Definition. (Equilibrium) Given parameters $\{r, v, \kappa, \tau_A, N, \mu, \bar{\eta}, g(\cdot), \bar{\omega}, \beta\}$ and model geography $\{\theta, \omega, \chi, T, H, T_F, \tau_F\}$, an equilibrium is a vector $\{b, \tau, \pi, \lambda, P^L, Q, L\}$ such that (i) associations maximize profits according to (5); (ii) the Nash bargaining condition (6) holds; (iii) commuter demand satisfies (8); (iv) passenger arrivals in (9) equal demand; and (v) bus loading probabilities in (1), (vi) queueing times in (2), as well as (vii) loading times in (3) are consistent with the queueing technology.

Market Mohring Effects

The model clarifies how the minibus loading process produces the increasing returns in queueing and loading times seen in Section III. I show in Appendix A.2 that the aforementioned *Indirect*

⁸By the usual properties of the Gumbel distribution, $\bar{W}^g \equiv v \log \left[\sum_{i,j,m} \exp \left[\theta_i^g + U_{ijm}^g + \omega_j^g \right]^{1/v} \right]$.

Market Mohring Effect of supply on total wait times in fact comprises two components:

$$\frac{\partial (Q_{ij} + L_{ij})}{\partial \log b_{ij}} \approx (1 - P_{ij}^L) \varepsilon_{g,b} \underbrace{[-\mu Q_{ij}^2 P_{ij}^L]}_{\text{Queueing}} + \underbrace{L_{ij}}_{\text{Loading}}. \quad (10)$$

Congestion in bus arrivals to a single loading area, as captured by the function $g(\cdot)$, precludes nonstop bus loading. Thus, greater supply reduces the amount of time during which the loading bay remains empty and thus shortens queueing times; higher baseline queueing times increase the magnitude of this first term. At the same time, when bus loading probabilities rise, buses fill more slowly, but this second term matters little in practice. The net effect grows in magnitude with the elasticity $\varepsilon_{g,b}$ of the bus arrivals function $g(\cdot)$ with respect to supply.

Similarly, passengers who board buses from a physical queue generate an almost mechanical demand externality. This *Direct Market Mohring Effect* of demand on total wait times,

$$\frac{\partial (Q_{ij} + L_{ij})}{\partial \log \lambda_{ij}} \approx \underbrace{-L_{ij}}_{\text{Loading}} + \underbrace{\lambda_{ij} Q_{ij}^2}_{\text{Queueing}}, \quad (11)$$

operates, in the first term, via loading times. Faster passenger arrivals mechanically fill buses faster and generate larger time savings, the longer baseline loading times and thus the larger bus capacity $\bar{\eta}$. Demand also directly increases queueing times; high queueing efficiency μ decreases the queueing time in the second term and thus dampens this spillover. Thus, efficient queueing and large buses beget negative Indirect and Direct Market Mohring Effects and thus increasing returns in the minibus queueing technology.

Efficiency

To consider when the market internalizes such spillovers, I first define welfare and the social planner problem.

Definition. (Welfare) Welfare, Ω , equals the aggregate ex-ante expected utility of newly-born commuters plus aggregate minibus profits $\Pi \equiv \sum_{i,j} \Pi_{ij}$, rebated to commuters upon birth.

$$\Omega \equiv \sum_g N^g \bar{W}^g + \Pi \quad (12)$$

The social planner then chooses allocations subject to the queueing technology.

Definition. (Planning Problem) The social planner chooses bus supply b for each route as well as skill-group-specific commute choice probabilities π for each home, work, and mode to maximize

welfare, Ω , as in (12), subject to the queueing technology (1)-(3) and (9) as well as the constraint $\sum_{i,j,m} \pi_{ijm}^g = 1$ for each skill group g .

The two Mohring Effects, themselves inherent to the physical minibus loading process, turn out to correspond to the two sources of inefficiency in the model, which I formally define as follows.

Definition. (Indirect Market Mohring Effect) Given an equilibrium $\{b, \tau, \pi, \lambda, P^L, Q, L\}$, associations internalize the Indirect Market Mohring Effect whenever b maximizes welfare, $\Omega(\cdot, \pi)$, conditional on commuter choice probabilities and subject to (1)-(3) and (9).

Definition. (Direct Market Mohring Effect) Given an equilibrium $\{b, \tau, \pi, \lambda, P^L, Q, L\}$, commuters internalize the Direct Market Mohring Effect whenever demand π maximizes welfare, $\Omega(b, \cdot)$, conditional on bus supply and subject to (1)-(3) and (9) as well as the adding-up constraint.

However, bargained fares correct neither the Indirect nor the Direct Market Mohring Effects, except in a double knife-edge case, in the spirit of Hosios (1990).

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

$$\left[\tau_{ij} - \frac{\chi_{ij}}{\bar{\eta}} - \frac{\bar{\omega}}{\bar{\eta}} (L_{ij} + T_{ij}) - \frac{\lambda_{ij} \bar{\omega}}{\bar{\eta}} \frac{\partial L_{ij}}{\partial \lambda_{ij}} \right] \frac{\partial \lambda_{ij}}{\partial b_{ij}} = - \left(\sum_{g'} N^{g'} \pi_{ijM}^{g'} r \omega_j^{g'} \right) \left[\frac{\partial Q_{ij}}{\partial b_{ij}} + \frac{\partial L_{ij}}{\partial b_{ij}} \right]. \quad (13)$$

Commuters internalize the Direct Market Mohring Effect if and only if

$$\tau_{ij} = \frac{\chi_{ij}}{\bar{\eta}} + \frac{\bar{\omega}}{\bar{\eta}} (L_{ij} + T_{ij}) + \sum_{g'} N^{g'} \pi_{ijM}^{g'} \left[r \omega_j^{g'} \left(\frac{\partial Q_{ij}}{\partial \lambda_{ij}} + \frac{\partial L_{ij}}{\partial \lambda_{ij}} \right) + \frac{\bar{\omega}}{\bar{\eta}} \frac{\partial L_{ij}}{\partial \lambda_{ij}} \right]. \quad (14)$$

Proof. See Appendix A.3. □

Associations choose efficient supply only if the marginal profits from increased demand, on the left-hand side of (13), equal passengers' value of the Indirect-Mohring-induced wait time savings. The greater these savings in (10), i.e. the greater the elasticity $\varepsilon_{g,b}$ and thus the less susceptible these bus arrivals to congestion, the more acutely associations under-provide minibuses. Commuters, in turn, make efficient choices only when fares exactly offset the value of the marginal Direct-Mohring gains. The latter, on the right-hand side of (14), comprise operations and labor costs plus the value of wait time saved. Thus, high queueing efficiency and large buses go hand-in-hand with inefficiently low demand levels. Generically, however, the physical and infrastructural constraints inherent to transit, as well as many other public goods, produce increasing returns and preclude efficient private provision.

TABLE 1. CALIBRATED PARAMETERS

Parameter	Description	Value	Parameter	Description	Value			
<i>Externally Calibrated</i>								
I	Number Locations	18	r	Commuter Rate of Time Pref.	0.001			
N^g	Commuter Populations		v	Gumbel Shape	4.76			
T_{ij}	Road Travel Time		κ_M^l	Low-Skill Minibus Util. Cost	7.7			
H_{ij}	Formal Wait Time		κ_M^h	High-Skill Minibus Util. Cost	15			
T_{ijF}	Formal Travel Time		κ_F^l	Low-Skill Formal Util. Cost	3.6			
τ_A	Car Commute Cost	5.2	κ_F^h	High-Skill Formal Util. Cost	9.2			
τ_{ijF}	Formal Fare		<i>Internally Calibrated</i>					
$\bar{\eta}$	Minibus Capacity	15	β	Assoc. Bargaining Power	0.09			
χ	Minibus Operating Cost	0.06	$\bar{\omega}$	Driver Wage	0.001			
<i>Minibus Supply</i>								
μ	Queueing Efficiency	11.82	<i>Model Inversion</i>					
α_0	Arrival Intercept	0.24	θ_i^g	Amenities				
α_1	Arrival Elasticity	0.38	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 supply parameters are estimated using the station counts. 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.

V. ESTIMATION OF MINIBUS AND DEMAND PARAMETERS

I estimate the model’s structural parameters in five main steps. First, I devise an instrumental variables strategy that employs the station counts of passengers and buses to identify the queueing efficiency μ . Second, from commuters’ stated preferences, I estimate their mode-specific utility costs κ_m^g , rate of time preference r , and Gumbel shock scale v . Third, I estimate the minibus arrival function $g()$. Fourth, I externally calibrate a geography composed of $I = 18$ *transport analysis zones* in Cape Town as well as other secondary parameters. Finally, conditional on all other parameters, I internally calibrate the association bargaining power β and driver wage $\bar{\omega}$. As part of this final step, I also 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.

Queueing Efficiency

First, I devise a new estimation methodology for queueing efficiency, μ , which leverages variation in passenger arrivals, dwell times, and passenger queues over time within a given minibus route. In contrast, existing methods typically require observations of “service”—in my case, boarding—times

(Asanjarani et al. (2021)) or nuanced sampling strategies difficult to implement in non-virtual queues (Chowdhury and Mukherjee (2011, 2013)). I estimate a linearized equation, derived in Appendix A.4 based on Equation (2), for the expected passengers in the queue for route ij during hour t , $n_{ijt} \equiv \frac{Q_{ijt}}{\lambda_{ij}}$. This queue length decreases with the fraction of time, $\text{Pr}(\text{bus dwell})_{ijt}$, during which a bus dwells in the loading bay, per arriving passenger, more so, the higher queueing efficiency, μ :

$$\log n_{ijt} = -\mu \frac{\text{Pr}(\text{bus dwell})_{ijt}}{\lambda_{ijt} \bar{\eta}_{ijt}} + h(\bar{\eta}_{ijt}) + \underbrace{\delta_{ij} + \delta_{p(t)} + \zeta_{ijt}}_{\equiv \varepsilon_{ijt}}. \quad (15)$$

Queue length also depends on a function $h(\cdot)$ of bus capacity as well as an unobservable term ε_{ijt} , which contains the rate ι_{ij} at which loading delays end.⁹

Such interruptions of the usual loading process, in ε_{ijt} , pose the primary threat to identification, which I neutralize with fixed effects and an instrument. Delays might directly lengthen queues as well as dwell times and thus bias μ towards zero. Route fixed effects, δ_{ij} , however, capture time-invariant causes, such as cramped stations, and time-of-day fixed effects, $\delta_{p(t)}$, absorb any higher delay-susceptibility of the busiest commute times. Thus, the error term ζ_{ijt} contains only idiosyncratic factors that might delay or stop loading: poor weather or special events, for example. I neutralize any remaining endogeneity with an instrument for dwell times per passenger: the number of neighborhood- i residents who leave for work during hour t , measured in a 2013 survey.¹⁰

This distribution of work start times instrument satisfies the exclusion restriction, provided it does not systematically vary with idiosyncratic interruptions to loading. Since I measure dwell times and passenger arrivals in 2022, such hour-by-hour trends in interruptions would have to persist over nine years to violate instrumental exogeneity. Even so, commute start times remain exogenous, provided the local timing of weather or events does not systematically vary by work start time. In Online Appendix D.1.1, I search for evidence of such systematic hourly trends in weather and event shocks, which would likely go hand-in-hand with a higher variance of observable indicators at the corresponding times of day. However, the variances of wind speed across days and traffic speed across individual vehicles do not appreciably differ by hour.

Table 2 displays the queueing efficiency estimated from (15) by instrumental variables. I start with a baseline specification in Column 1; in Columns 2 and 3, I add route and time-of-day fixed effects, respectively. The first-stage F statistics reveal a borderline-weak instruments problem, so I

⁹I calculate variables on the right-hand side of (15) from route-by-hour averages of bus dwell probabilities, passenger arrivals per minute, and bus occupancy. Neither the dependent nor the independent variable contain any zeros.

¹⁰At the time of writing, the 2013 Cape Town Household Travel Survey offered the most recent available spatially-granular data on commuter demand.

TABLE 2. QUEUEING EFFICIENCY ESTIMATES

Parameter	(1) log queueing passengers	(2) log queueing passengers	(3) log queueing passengers
μ	10.70	11.82	11.27
<i>Queueing Efficiency</i>	(7.61)	(4.07)	(4.15)
<i>Weak IV-Robust P-Value for $\mu = 0$:</i>	0.097	0.002	0.001
Route FE		✓	✓
Time-of-day FE			✓
Observations	121	117	117
First-Stage F Statistic	2.12	5.75	6.42

Notes: Robust standard errors in parentheses, clustered at the origin level. Table presents estimates of (15) over hours and 44 routes in my station count data, with fixed effects included, as noted, for route and time of day (mid-peak, 7-9am, vs. edge of peak, 6-7 or 9-10am). I instrument for the bus dwell probability per arriving passenger measure using the log number of commuters living in the mesozone spatial unit where the route originates who report leaving their home during hour t , calculated from the 2013 Cape Town Household Travel Survey, and drop one outlier observation. I additionally report the effective first-stage F statistic following Olea and Pflueger (2013) as well as the p-value associated with the Anderson-Rubin chi-squared test robust to weak instruments (Andrews et al. (2019)).

confirm that the estimated parameters remain significant under an Anderson-Rubin test robust to weak instruments. The efficiency $\hat{\mu} = 11.82$ in my preferred, most precisely estimated specification in Column 2 indicates that the average commuter takes about 5 seconds to board a minibus. The literature offers little precedent for comparison, but this short boarding time suggests that additional minibus demand would only minimally increase queueing times.

Stated Preferences

Second, the stated preference surveys' exogenous variation in commute attributes permits direct estimation of the model's logit demand system via standard maximum likelihood. The attributes in the survey correspond to the model in Section IV, with one notable addition: each alternative l featured a set $\mathcal{Z}(l)$ 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, which would include queueing and loading, 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

$$\pi_{il}^g = \frac{\exp \left[-\kappa_{m(l)}^g - \sum_{z \in \mathcal{Z}(l)} \xi_z^g - r\omega_i(H_l + T_l) - \tau_l + rH_l\tau_l \right]^{1/\nu}}{\sum_{l'} \exp \left[-\kappa_{m(l')}^g - \sum_{z \in \mathcal{Z}(l')} \xi_z^g - r\omega_i(H_{l'} + T_{l'}) - \tau_{l'} + rH_{l'}\tau_{l'} \right]^{1/\nu}}. \quad (16)$$

¹¹Unlike 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.3.

Differences in mode shares, all other attributes equal, identify relative utility costs, κ_m^g .¹² 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 v and allows me to calculate dollar willingness to pay for each attribute.

The left panel of Table 3 provides developing-country estimates of fundamental travel demand parameters, as estimated from (16). The rate of time preference, $\hat{r} = 0.001$, implies a value of time, 24% of the hourly wage, at the lower end 20-140% typical for the developed countries previously studied (Small and Verhoef (2007), Almagro et al. (2024), Buchholz et al. (2024), and Goldszmidt et al. (2020)). The substantially lower value in this sub-Saharan African context, even relative to existing developing-country estimates for India (Kreindler (2022)), further justifies policies aimed at non-time barriers such as security. Next, I estimate a Gumbel scale parameter $\hat{v} = 4.76$; this high idiosyncratic preference variance and the low status quo fares generate a minibus own-price elasticity of demand of -0.17 . In contrast, Goldszmidt et al. (2020) and Almagro et al. (2024) estimate significantly higher own-price elasticities of -0.5 to -0.7 for public transit or ride-hailing in the US. In consequence, changes in fares alone might not induce sizeable enough mode shifts to correct the Market Mohring Effects. However, Cape Town drivers, with an implied own-price elasticity of -1.62 , substitute readily towards other modes.¹³

The remaining parameters in Table 3 speak to the *quality* of privatized shared transit and specific deficiencies which policymakers might target. The low- and high-skill minibus utility costs, $\hat{\kappa}_M^l = 7.68$ and $\hat{\kappa}_M^h = 15.03$, demonstrate that both groups dislike minibuses relative to formal transit and cars, given equal cost and travel time. In contrast, commuters in Latin America, by the same measure, prefer minibuses to formal bus rapid transit (Tsivanidis (2023) and Zarate (2024)). I next unpack the root causes, from crime risk to discomfort, of this perhaps atypical distaste for minibuses in South Africa.

Indeed, all three quality improvements which I test in the right panel of Table 3 significantly decrease minibuses' utility cost for the low-skill group. Most strikingly, the high-skill would pay a full \$2.75 per commute for station security guards—whose presence might deter muggings and harassment. This value, scaled up to an annual level, exceeds women's already-sizeable willingness to pay for safer walking routes in India in Borker (2021) by up to a factor of five. Security concerns,

¹²While 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.

¹³To calculate demand elasticities, I use median minibus fares from my on-board tracking data as well as the median calibrated formal fare and calibrated car per-commute cost. Mode shares come from the 2013 Cape Town Household Travel Survey.

TABLE 3. STATED PREFERENCE SURVEY ESTIMATES

Parameter	Value		Parameter	Value	
	Low-Skill	High-Skill		Low-Skill	High-Skill
r	.001		ξ_{security}	-1.09	-2.75
<i>Commuter Rate of Time Pref.</i>	(.0004)		<i>Station Security</i>	(0.390)	(0.84)
v	4.76		$\xi_{\text{no overloading}}$	-1.38	-1.39
<i>Gumbel Scale</i>	(1.26)		<i>Overloading Ban</i>	(0.437)	(0.543)
<i>Utility Costs</i>	<i>Low-Skill</i>	<i>High-Skill</i>	$\xi_{\text{follows speed limit}}$	-1.36	-0.825
κ_F	3.63	9.17	<i>Speed Limit Enforcement</i>	(0.445)	(0.465)
<i>Formal Transit Utility Cost</i>	(0.51)	(1.89)			
κ_M	7.68	15.03			
<i>Minibus Utility Cost</i>	(1.56)	(3.55)			

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 (16). I normalize $\kappa_A^g = 0$ and restrict the sample to individuals employed outside the home between 25 and 65 years of age.

though certainly not exclusive to minibus stations in South Africa, evidently underlie a sizeable part of commuters' distaste for minibus travel.

Lastly, I provide evidence that, as required for my policy counterfactuals, these preferences represent the entire Cape Town commuter population. Recall that the stated preference samples mirror the population along every demographic dimension except, in the case of my own survey, actual commute modes. I re-estimate the logit model (16) and weight my survey by realized commute mode choices. Separately, I restrict the sample to respondents interviewed outside of minibus stations. The results, in Online Appendix D.1.2, change little, so the estimated values of time and quality would not appear to reflect those of only a selected subgroup of minibus users.

Minibus Arrivals

Third, I estimate a log-linear function, $g(b_{ij}) \equiv \alpha_0 b_{ij}^{\alpha_1}$, for minibus arrivals. By the properties of Poisson processes, the gap in minutes between the departure of one bus from the origin station and the arrival of the next averages $1/\lambda_{ij}^B \equiv 1/(g(b_{ij}) \xi_{ij})$. Thus, the average $\overline{\text{bus gap}}_{ijt}$ observed in the station counts decreases with the number of buses b_{ijt} on a route ij during hour t according to $\log(\overline{\text{bus gap}}_{ijt}) = -\log \alpha_0 - \alpha_1 \log b_{ijt} + \vartheta_{ijt}$. The unobserved arrivals efficiency, $\vartheta_{ijt} \equiv -\log \xi_{ij}$, chiefly reflects overall foot and car traffic levels in origin neighborhoods, certainly correlated with associations' chosen bus supply. Thus, in Online Appendix D.1.3, I additionally include origin-hour fixed effects and instrument for bus supply with origin-destination distance. Route length determines

operations costs and thus supply but should not systematically vary with the degree to which, say, surrounding foot traffic affects specific routes at the same station.

In the baseline OLS specification, I find an intercept of $\hat{\alpha}_0 = 0.24$ and an arrival elasticity of $\hat{\alpha}_1 = 0.38$.¹⁴ The latter falls within the confidence intervals of the aforementioned fixed-effects and instrumental variables estimates. The implied congestion index $1 - \hat{\alpha}_1$, relative to the on-road congestion elasticities estimated by Allen and Arkolakis (2022), implies that minibus stations congest somewhat more readily.¹⁵ Greater susceptibility to congestion, in turn, constrains the extent to which the Indirect Market Mohring Effect can ameliorate queueing times.

Calibration and Inversion

Fourth, I directly obtain the car commute cost τ_A , minibus capacity $\bar{\eta}$, and the model geography from the data. The latter includes commute populations N^g , minibus operating costs χ_{ij} , travel times $\{T_{ij}, T_{ijF}\}$, and formal transit wait times H_{ij} as well as fares τ_{ijF} .¹⁶ I provide additional details regarding all externally calibrated parameters in Online Appendix D.2.

Finally, I match aggregate moments in my newly-collected minibus data to obtain the remaining supply parameters. For each parameter iteration, I 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_{ijt} , across routes and five-minute time 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.¹⁷

The second and third columns of Table 4 list the values of each moment in the data and in the calibrated model across routes. Heuristically, the median fare identifies association bargaining power β , and queue length identifies the minibus driver opportunity cost $\bar{\omega}$. The final column lists the calibrated parameter values. Associations' market power looms large in the public imagination, yet their minuscule estimated bargaining power reveals that status quo fares actually fall far short of monopoly levels.

¹⁴ $\hat{\alpha}_0$: standard error = 0.10; $\hat{\alpha}_1$: standard error (clustered by origin) = 0.16, $N = 161$ routes-by-hours, $R^2 = 0.088$. I present full results in Online Appendix D.1.3 and employ the baseline OLS estimates in the model quantification to obtain a transparent single estimate of α_0 .

¹⁵The index $1 - \hat{\alpha}_1$ is the correct measure of congestion under the assumption that absent traffic jams, bus supply would reduce the time between loading buses with an elasticity of unity.

¹⁶Since the fixed 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.2. I calculate formal transit fares τ_{ijF} directly from Cape Town's distance-based public transit fare scheme.

¹⁷I 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

TABLE 4. INTERNAL CALIBRATION

Moment			Parameter		
Description	Data	Model	Description		Value
Median Minibus Fare	1.05	1.05	β	Association Bargaining Power	0.09
Median Queue Length	4	4	$\bar{\omega}$	Minibus Driver Wage	0.001

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. For calibration, I choose a close-to-optimal starting point, which I then feed into the simplex search method to numerically minimize the sum of squared (percentage) deviations from the two moments. In each iteration, I invert the model equations for employment by residence and workplace to obtain implied residential amenities and workplace wages.

VI. MODEL-PREDICTED MINIBUS OPERATIONS AND AGGREGATE PATTERNS

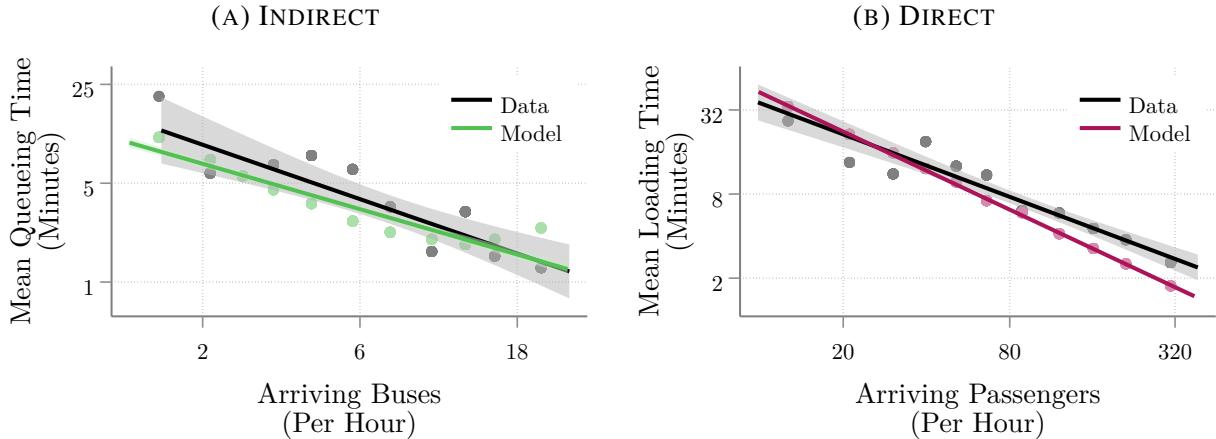
I now demonstrate that my model matches non-targeted data: the minibus network, the Market Mohring Effects, and the actual commute modes of stated preference respondents. First, the model-predicted bus supply on each origin-destination pair, in Online Appendix Figure A.2, mirrors the number of distinct minibus routes between each pair of neighborhoods.¹⁸ Importantly, the model matches the concentration of minibuses in central neighborhoods and the direct links between outlying suburbs. However, demand patterns and, thus, the relative magnitude of queueing versus loading times differ vastly across these two groups of routes.

Second, the model replicates the levels of queueing and loading times as well as the Indirect and Direct Market Mohring Effects. Figures 5a and 5b again feature queueing times versus arriving buses and loading times versus arriving passengers, respectively. I plot binned scatterplots of route-by-hour observations in the station count data in black and routes in the model in color. Cape Town's long *passenger* queues, in both the data and model, contrast starkly with other developing-country cities (Cervero and Golub (2007)) where, instead, "vehicles spend most of their day waiting for their turn" (Kerzhner (2023)). Nairobi, for example, evidently allows for higher bargained minibus fares or benefits from faster levels α_0 of bus arrivals. Furthermore, both the queueing times in Figure 5a and the loading times in Figure 5b fall in a Market-Mohring fashion with route scale, precisely as in the data. The estimated bus congestion elasticity α_1 , queueing efficiency μ , and bus size $\bar{\eta}$ thus jointly deliver a successful approximation of the actual minibus loading process.

skill-group average.

¹⁸Note that, in reality, unlike in my model, many neighborhood (transport analysis zone) pairs are linked by multiple distinct minibus routes.

FIGURE 5. MARKET MOHRING EFFECTS IN DATA VERSUS MODEL



Notes: Panel (A) displays binned scatterplots and best-fit lines of the log-scale relationship between expected passenger queueing time, Q_{ij} , and newly-arriving buses per hour, proportional to $\lambda_{ij}^B (1 - P_{ij}^L)$, 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 minibus loading times, L_{ij} , and newly-arriving passengers per hour, proportional to λ_{ij} .

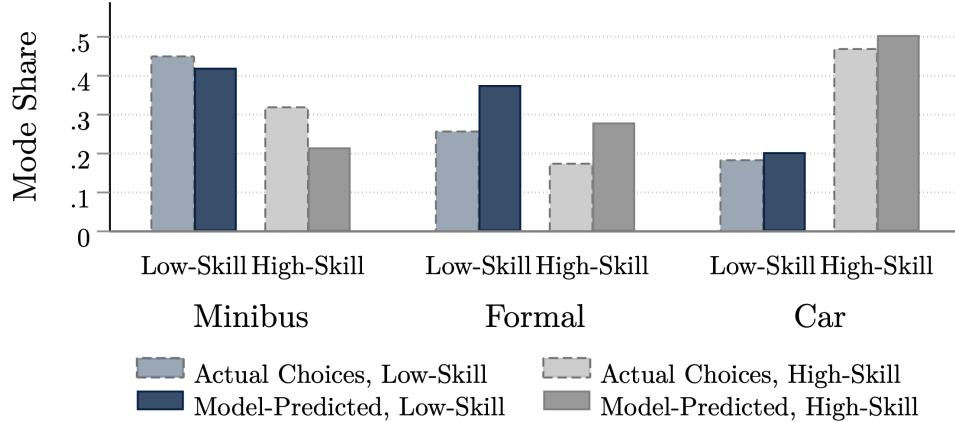
Third, the model accurately predicts the actual reported commute choices of stated preference respondents. The bars in Figure 6 indicate the skill-group-level shares of respondents in the combined stated preference sample who commute via each mode in reality (light colors) and as predicted by the model (dark colors). Respondents' actual *revealed* commute choices mirror model predictions based on their hypothetical *stated* choices. In particular, the model, chiefly via utility costs, replicates the skill differential in minibus use as well as the fact that the high-skill overwhelmingly drive. Thus, the stated preference methodology appears to accurately capture real-world preferences.

Because my later results depend crucially on the associated estimates, I demonstrate the plausibility of respondents' stated preferences in two additional ways. First, in Online Appendix E.1, I show that the model matches origin-destination-level mode shares. Second, in Online Appendix E.2, I confirm that demographic heterogeneity in commuters' values of time and quality improvements largely follows intuition. Women, for example, place a higher value on time saved, as Borghorst et al. (2021) similarly find.

VII. URBAN TRANSPORTATION POLICIES

Finally, I use the estimated model to quantify the gains from policies that leverage Cape Town's existing minibus system. First, I investigate the implementation of the social planner's optimal minibus supply and demand via bus and commuter subsidies. Second, I quantify the fraction of

FIGURE 6. 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_{ijm}^g ; the model predictions are sums of these choice probabilities within each skill and mode.

the associated welfare gains that more straightforward market structure changes, such as monopoly pricing or free entry, can likewise attain. Third, I calculate the gains from upgrades to the minibus technology: expanded stations, smaller buses, and the provision of security at the publicly-owned minibus stations. I present changes in the welfare measure Ω defined in (12) as equivalent variation: the proportionate change in a skill group's wages ω_j^g , at baseline values of $\{b, \tau, Q, L, \kappa, \bar{\eta}\}$, that leaves the average commuter equally well off as in the counterfactual.¹⁹ Table 5 at the end of this section summarizes all counterfactuals.

Social Planner Optimum

First, I implement the socially optimal minibus entry b and commuter choices π via subsidies to associations and minibus commuters.²⁰ Figure 7a displays the optimal commuter inflow λ_{ij}^* , relative

¹⁹To calculate welfare at the skill-group level in a manner unaffected by minibus fares and transfers, 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). I allocate the monetary costs of the security guard counterfactual according to population shares.

²⁰In Online Appendix C.2, I derive the per-minibus-commuter subsidies t_{ij} and per-supplied-minibus subsidy z_{ij} which, under free association pricing, $\beta = 1$, induce the socially-optimal commuter choices π and bus supply b . Commuters of a given skill pay equal lump-sum taxes to fund the total commuter subsidies paid out to their own

to the status quo, versus a critical determinant of demand: the wages ω_j^g at a route's destination. Notably, passengers queue efficiently enough that higher demand increases queueing times less than it decreases the high loading times of the relatively large Cape Town minibuses. Thus, the planner leverages the negative Direct Market Mohring Effect of demand on total wait times, as in (11), and optimal exceeds status-quo demand across all wage levels. Figure 7b instead plots the optimal minibus supply relative to the status quo on the vertical axis. Though more readily congestible than on-road trips, bus arrivals increase sufficiently quickly with supply that the planner would leverage the Indirect Market Mohring Effect in (10) and add buses to all routes.

Inefficiently-low demand and minibus under-provision, however, most acutely plague opposite sets of routes. In particular, low-wage, low-demand routes suffer the longest loading times, so the planner disproportionately increases their demand levels. The routes with the largest demand increases, mapped in Figure 7c, tend to connect far-flung suburbs, where previously thin demand filled buses inefficiently slowly. Logically, high-wage, high-demand routes have the opposite problem: buses fill quickly, but, conditional supply, passengers wait in longer queues to board in the first place. Thus, the planner disproportionately boosts bus supply on these busy routes which, in Figure 7d, connect major townships such as Khayelitsha and Mitchell's Plain to the CBD.

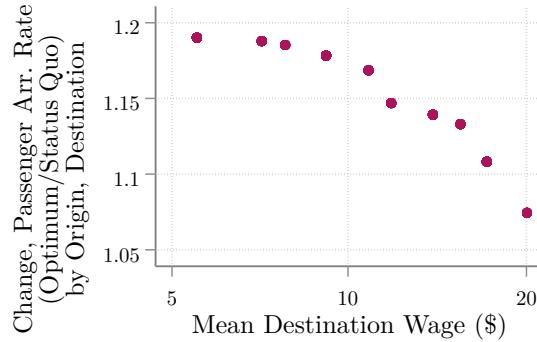
The optimization of the minibus system alleviates wait times everywhere but particularly to lower-income workplaces. The high-demand routes, which benefit from disproportionate bus supply increases, already enjoyed the lowest loading and queueing times, the latter thanks to the Indirect Market Mohring Effect. These high-wage routes' queueing and loading times, in Figure 8a, thus decrease little in absolute terms. In contrast, low-wage routes previously suffered the longest loading times, benefit from disproportionate demand increases, and thus experience the largest absolute declines in both forms of wait time. The planner-induced changes in work location choice probabilities, analogously plotted in Figure 8b, reveal slight spatial reallocation towards lower-wage workplaces. These growing commute flows, mapped in Figure 8c, span the aforementioned suburban links and radial trunk routes. Consequently, lower-income workers benefit disproportionately, but, gross of commute costs, average earned wages, in Table 5, scarcely change.

Nonetheless, the commute time gains outweigh these shifts towards lower-wage locations. As a result, low-skill commuters gain 1.35% in equivalent variation from the social optimum, relative to the status quo, and the high-skill gain 0.4%. Table 5 also spells out the extent of the mode shift from cars to *shared* minibuses. Back-of-the-envelope valuations of the implied reductions in greenhouse gas emissions, as detailed in Online Appendix D.2.6, increase the gains by another

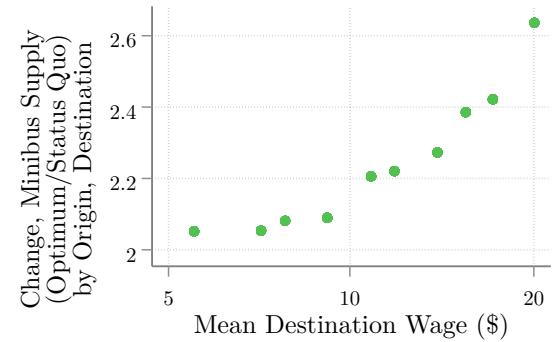
skill group as well as a share of minibus supply subsidies equal to their aggregate population share. In practice, Cape Town might leverage existing cash transfer programs, regulatory relationships with associations, and mobile-app-based check-ins to distribute such incentives.

FIGURE 7. SOCIAL PLANNER OPTIMUM

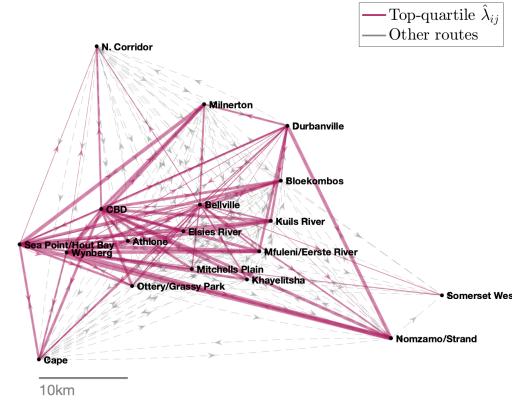
(A) OPTIMAL MINIBUS DEMAND VS. WAGES



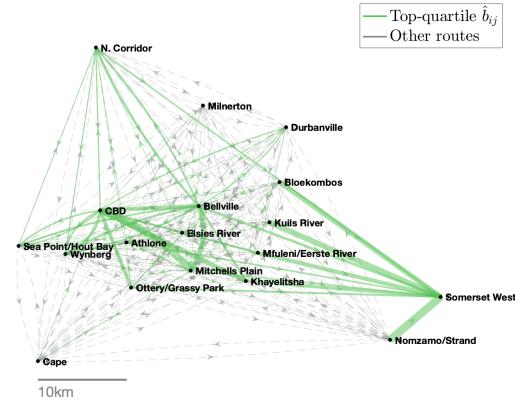
(B) OPTIMAL MINIBUS SUPPLY VS. WAGES



(C) LARGEST DEMAND INCREASES



(D) LARGEST SUPPLY INCREASES



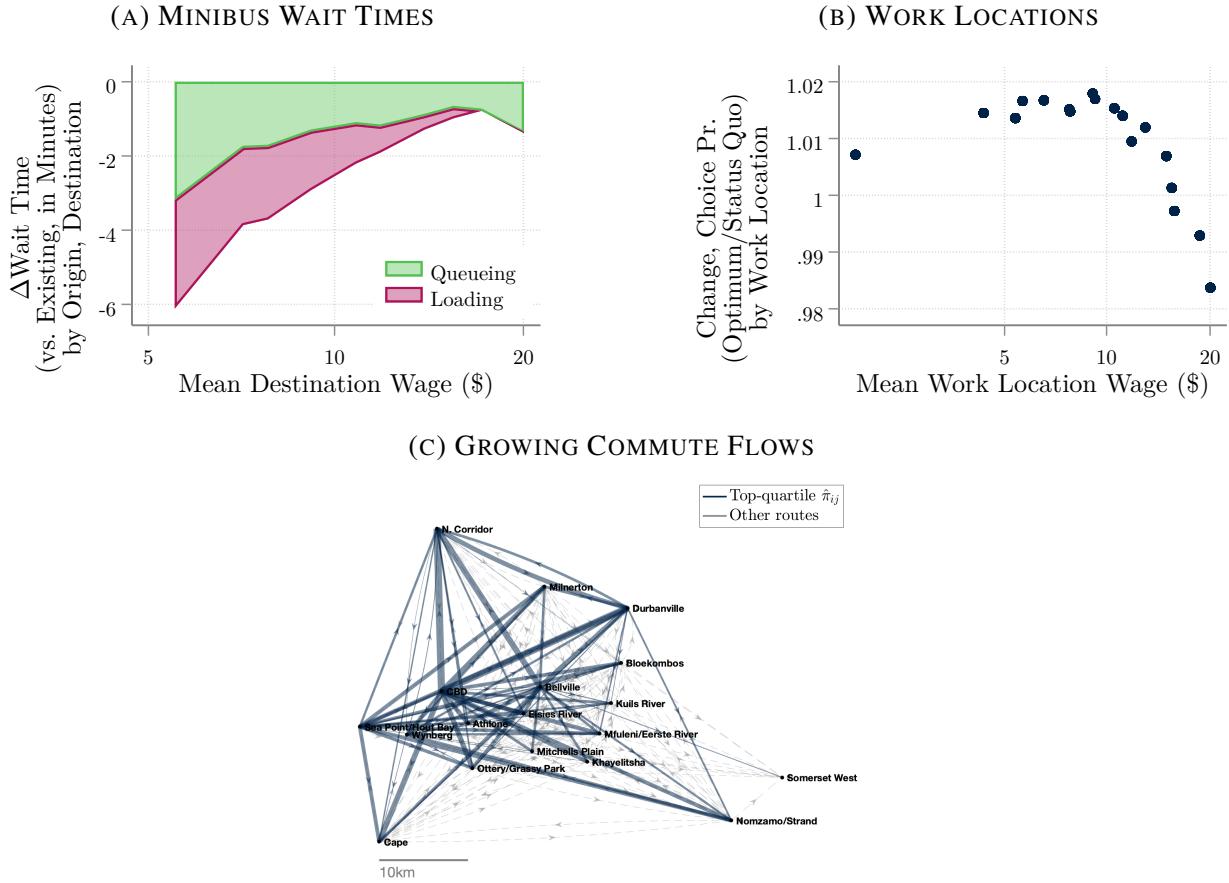
Notes: Panels (A) and (B) display, on the vertical axis, the social planner's optimal commuter inflow λ_{ij}^* and optimal minibus supply b_{ij}^* , respectively, in changes relative to the status quo at the route level. The horizontal axis displays the average across skill groups, weighted by aggregate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

2-5%. Thus, though workers, on average, commute to lower wages, the social optimum boosts aggregate welfare, reduces emissions, and improves equity within and across skill groups. These gains remain qualitatively similar when I extend the model in Online Appendix E.3 to incorporate (i) on-road congestion, (ii) nested logit demand, and (iii) endogenous bus departure timing.

Market Structure

Might simpler, more feasible changes to the minibus market structure approximate the social optimum? I first investigate the extent to which association monopoly pricing, i.e. full bargaining power $\beta = 1$, corrects the Market Mohring Effects. Removal of government constraints more than triples the minibus fares in Figure 9a. Associations take advantage of their market power and throttle

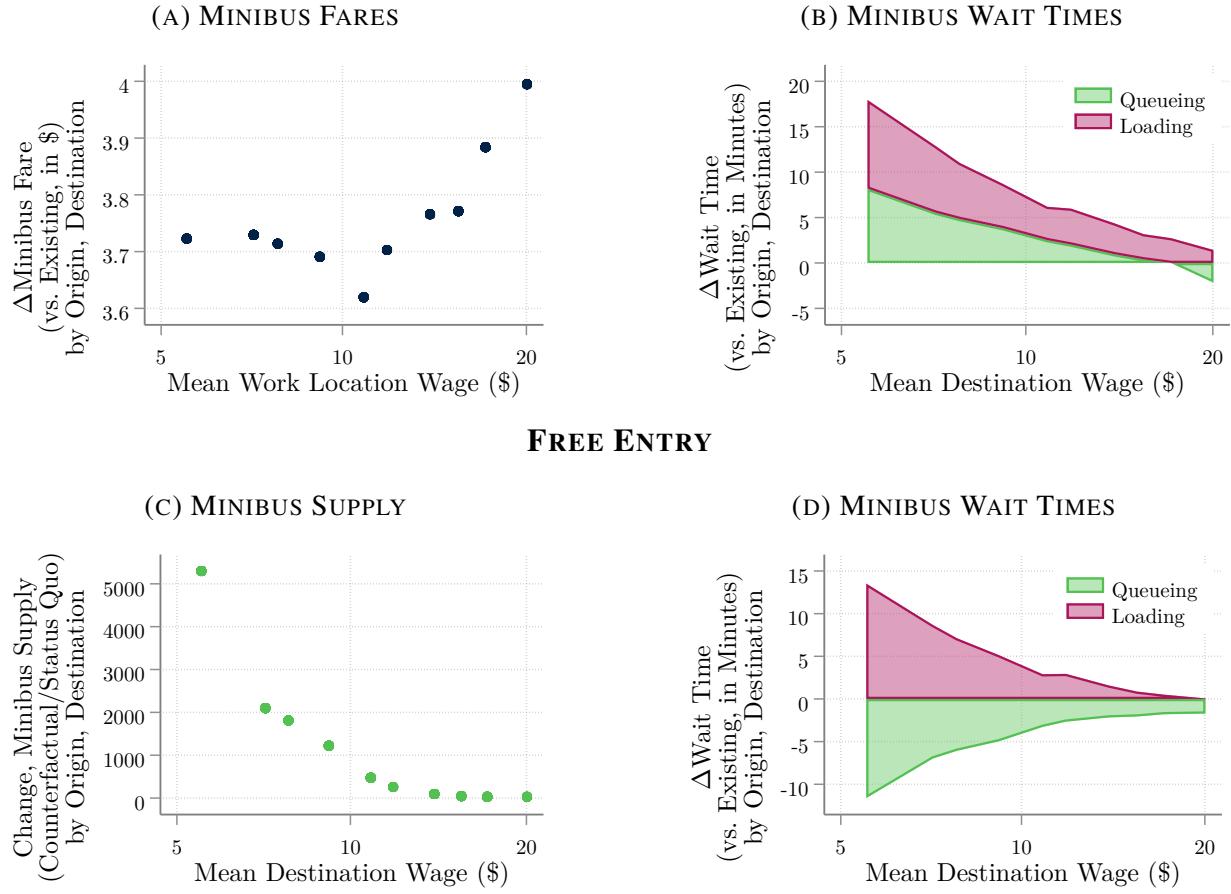
FIGURE 8. SOCIAL PLANNER: LOWER WAIT TIMES AND SPATIAL REALLOCATION



Notes: Panel (A) displays, on the horizontal axis, (route) destination wages, calculated as the average across skill groups, weighted by aggregate populations, and, on the vertical axis, the median route-level raw changes in queueing and loading times, Q_{ij} and L_{ij} , across routes with the corresponding destination wage. Panel (B)'s vertical axis instead displays a scatterplot of the proportionate change in work location choice probability, averaged over skill groups. Changes are calculated from the status quo to the social optimum. Panel (C) maps the 25% of home-work location pairs with the largest proportionate increases in choice probabilities, from the status quo to the social optimum. Line width indicates the magnitude by which social optimum exceeds status quo.

bus supply, though less so on the routes with the highest wages and thus willingness to pay, as in Online Appendix Figure A.3b. In consequence, queueing times—in Figure 9b—rise on all but the highest-wage routes. Slower passenger arrivals, shown in Online Appendix Figure A.3a, also push up loading times everywhere. Nevertheless, the aforementioned bias in favor of high-wage routes as well as the shift towards cars, in Table 5, allows commuters, in Online Appendix Figure A.3c, to reach higher-wage locations. However, the direct losses from longer waits dominate and particularly hurt low-skill commuters, who lose 1.5% in welfare terms and even more, net of induced carbon emissions. Thus, full association bargaining power pushes equilibrium further from the social optimum.

FIGURE 9. CHANGES IN MARKET STRUCTURE
MONOPOLY PRICING



Second, in light of the stark losses from monopoly pricing, I evaluate its opposite: free entry of minibuses. Specifically, as detailed in Appendix A.5, individual minibuses Nash-bargain fares with commuters, and free entry of minibuses drives profits to zero. Since I lack data on this hypothetical market structure, I agnostically allocate buses half of the bargaining power. Predictably, free entry drops fares, as in Online Appendix Figure A.3d, and bus supply, in Figure 9c, rises everywhere. Buses fill more slowly, and Figure 9d clarifies that higher loading times largely negate the time saved in shorter queues. Welfare, in Table 5, changes little, though emissions fall; free entry thus proves too blunt an instrument.

Technology

My last set of policy counterfactuals change the minibus “technology.” First, motivated by the queues in Figure 3b, I add a second independent bus loading bay and passenger queue for each route. Passengers randomly arrive to only one of the two queues, so buses fill more slowly—and loading times rise most on the already slow-to-fill low-wage routes.²¹ Thus, associations, in Online Appendix Figure A.4b, disproportionately withdraw buses from these low-demand routes, so not only loading but also queueing times rise, as in Figure 10a. I additionally plot the 30th percentile of the changes in wait times at each wage level as a dotted line in Figure 10a. Not surprisingly, a subset of very high-demand routes with high wages benefit from massive decreases in queueing time. Figure 10b demonstrates that commuters shift towards these high-wage work locations, and average wages rise substantially. Since precisely the busiest routes benefit from shorter queues, the net welfare gains rival those from the social planner’s optimization of the existing technology. A better-targeted policy – second queues only on the 10% of routes with the highest baseline demand λ_{ij} – could further boost these gains by 6-50%, as I discuss in Online Appendix E.3.

Second, a smaller bus size of $\bar{\eta}' = 12$ – in my model, equivalent to a mandate that buses depart with 12 passengers – would save loading time, both mechanically and, in Figure 10d, in equilibrium.²² Furthermore, the lower profitability of smaller buses allows associations to extract higher fares in Nash bargaining, as in Online Appendix Figure A.4c, so demand in Figure 10c and queueing times fall. Though the wait time gains rival those of the social planner optimum, higher fares limit the number of commuters who benefit and, more severely, the welfare effects.²³

Third, I evaluate the government provision of security guards at minibus stations.²⁴ I adjust the minibus utility cost κ_M^g by the estimated binary security effect ξ_{security}^g , and commuters pay their lump-sum share of guard wages.²⁵ Because they place a larger premium on security, high-skill commuters shift even more strongly towards minibuses than their low-skill counterparts. Buses fill faster, so associations boost supply (Online Appendix Figure A.4f) and thereby actually shorten

²¹I assume that bus arrivals to *each* queue are given by the same function $g(\cdot)$ of bus supply.

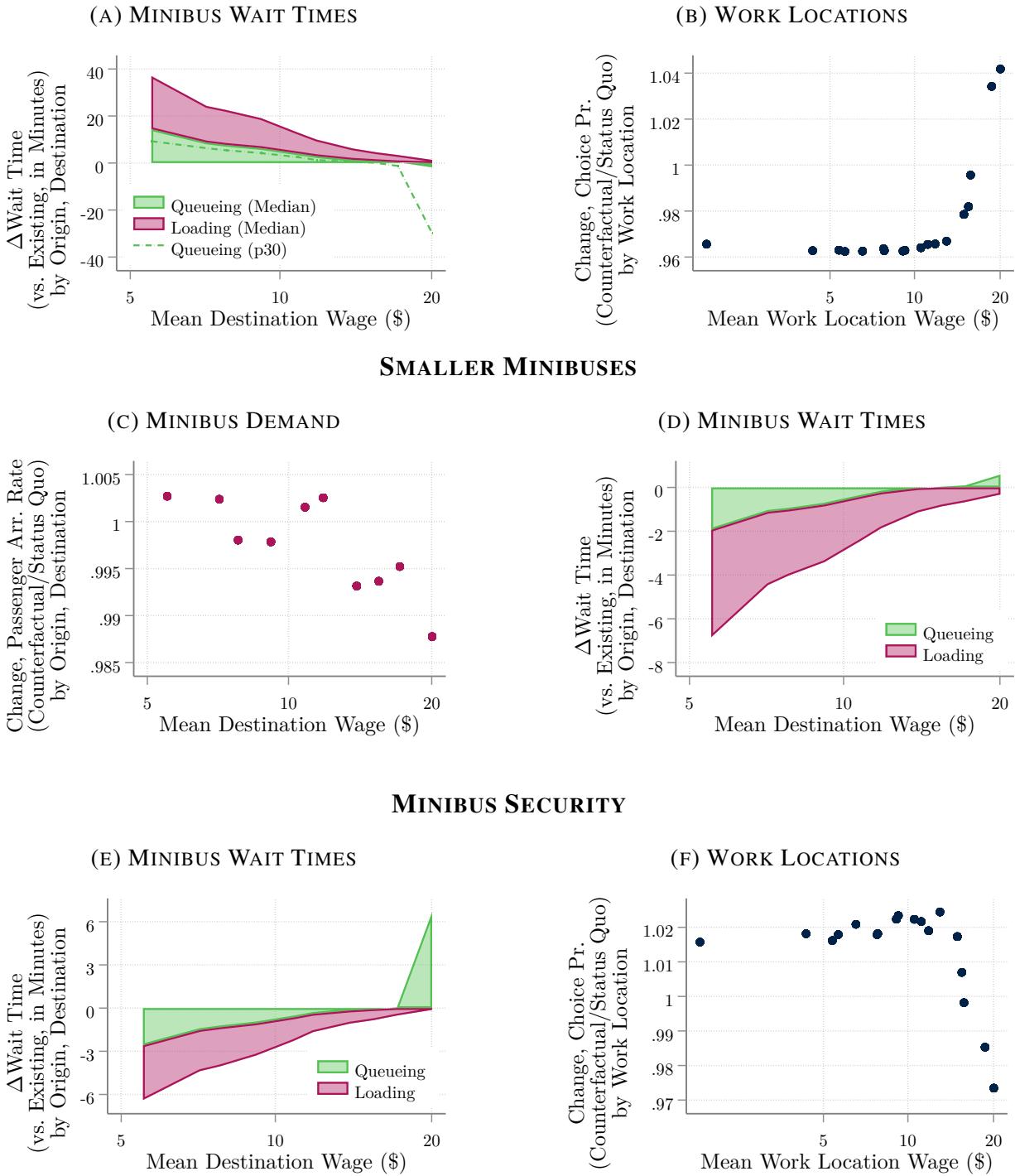
²²More precisely, bus size and a mandated threshold number $\bar{\eta}$ of passengers with which buses depart are isomorphic under the admittedly strong but not highly consequential assumption that operations costs χ_{ij} do not depend on bus size. Since I lack reliable operations cost data on hypothetical smaller minibuses, I do not adjust χ_{ij} in any counterfactual. Additionally, I confirm in Online Appendix E.4 that, in a model with endogenous bus departure thresholds, simply decreasing driver wages will not induce buses to depart with less than full loads.

²³I again test a more-targeted policy, in Online Appendix E.4, of smaller buses on only the 20% of routes with lowest baseline demand λ_{ij} , which, via similar mechanisms, essentially leaves welfare unchanged.

²⁴Anecdotal observation indicates that, in the status quo, only a tiny subset of stations employ security guards, so I calculate effects relative to a zero-guard baseline.

²⁵The hourly guard wage, at only twice the median minibus fare, plays a minuscule role in welfare. I assume two guards per route; commuters pay a lump-sum tax to cover four hours of guard costs during the morning peak commute at a wage quoted by a local security firm.

FIGURE 10. CHANGES IN TECHNOLOGY
TWO QUEUES



Notes: Figure displays outcomes under technology counterfactuals on the vertical axes, always versus a route's destination wage, averaged across skill groups and weighted by aggregate populations, on the horizontal axis. Panels (A), (D), and (E) display the median route-level raw changes in queueing and loading times, Q_{ij} and L_{ij} , across routes with the corresponding destination wage. Panel (C) displays a binned scatterplot of the change in commuter inflow λ_{ij} under a given technology counterfactual relative to the status quo. Panels (B) and (F) display binned scatterplots of the proportionate change in work location choice probability, averaged over skill groups, in a given counterfactual relative to the status quo.

TABLE 5. COUNTERFACTUAL URBAN TRANSPORTATION POLICIES

Policy	Skill:	Change in Mode Share				% Change in...					
		Minibus		Car		Earned Wage	Emissions	Welfare		Welfare, Net of Emissions	
		Low	High	Low	High			Low	High	Low	High
Social Planner		0.03	0.03	-0.01	-0.02	-0.33	-2.14	1.35	0.4	1.39	0.42
Monopoly Pricing		-0.11	-0.08	0.02	0.04	0.78	7.09	-1.47	-0.56	-3.09	-1.26
Free Entry		0.03	0.03	-0.01	-0.01	0.02	-2.14	-0.02	0	0.02	0.02
Two Queues		0.03	0.01	-0.01	-0.01	0.89	-1.63	1.6	0.24	1.63	0.25
Smaller Minibuses		0	0	0	0	-0.05	0.2	-0.24	-0.04	-0.24	-0.04
Minibus Security		0.04	0.09	-0.01	-0.05	-0.5	-4.08	2.35	2.34	2.41	2.36

Notes: Table summarizes the effects of all counterfactuals: implementation of the social optimum via optimal minibus and commuter subsidies, monopoly pricing ($\beta = 1$), free entry of minibuses, adding a second queue to each minibus station, reducing minibus capacity $\bar{\eta}$ to 12, and adding security guards to all minibus stations. The first four columns show the changes in the minibus and car mode shares by skill group. The fifth and sixth show the percent changes in the average wage *earned* by commuters, gross of commute costs, and total emissions, which I calculate as described in Online Appendix D.2.6. The final four columns show the percent change in group-level welfare, measured as equivalent variation and, in the last two columns, net of external emissions costs.

queues on most routes, as demonstrated by Figure 10e. However, on the highest-wage, highest-demand routes, buses already loaded without interruption in the status quo, so queueing times rise along with demand. In Figure 10f, commuters shift relatively away from the highest-wage work destinations, in contrast to the case of second queues. However, the utility and emissions gains far outweigh the slight decreases in average income, so the welfare benefits of this simple intervention, at around 2.4%, exceed those of any other policy.

VIII. CONCLUSION

In this paper, I build a model of the privatized shared transit sector which dominates many developing-country cities. My theory unpacks the literature's typically exogenous transport costs via a two-sided queueing model, which generates a market-driven version of the Mohring (1972) increasing returns in wait times. Newly-collected data on passenger and bus queues in Cape Town permits direct estimation of queueing efficiency and congestion in bus arrivals. Furthermore, I introduce the stated preference approach to the urban literature to identify the commuter demand system. Finally, I quantify the scope for improvement in the (in)efficiency of Cape Town's current privatized transit provision. I then compare the model-implied gains from the optimization of minibus supply and demand to the effects of potentially more feasible policies. The latter include changes in market structure as well as upgrades to the technology of the minibus system.

Three policies' welfare and equity gains stand out. The social planner harnesses the Indirect and Direct Market Mohring Effects to correct inefficiencies fundamental to physical queues and multi-passenger buses. The associated decreases in queueing and loading times generate welfare gains that rival or exceed those of commonly studied urban transit infrastructure. Technological upgrades, such as additional passenger queues or minibus station security guards, prove similarly beneficial, though only in the former case do workers sort towards higher-wage workplaces. All three policies decrease carbon emissions and disproportionately benefit low-skill commuters. Even absent the capacity for more substantial infrastructure investments, policymakers thus enjoy ample scope to improve upon the privatized provision of transit.

APPENDIX

A. THEORY

A.1 Wait Time Derivations

First, 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). 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, $\lambda \equiv \frac{\tilde{\lambda}_{ij}}{\bar{\eta}}$. 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. Both the (independent) bus arrival and the bus loading Poisson shocks must occur for the next bus to begin loading, so this interval between successive loading intervals equals the mean of the corresponding hypoexponential distribution, i.e. $E(t_r) \equiv \frac{1}{\lambda_{ij}^B} + \frac{1}{t_{ij}}$. Finally, Avi-Itzhak and Naor (1963)'s probability that the service station is in order equals the probability that a bus is loading, $p_0 \equiv P_{ij}^L$.

Then, in order to characterize wait times, I derive the probability that a bus is loading on a route, which must satisfy

$$P_{ij}^L = \frac{1}{1 + \frac{\tilde{\lambda}_{ij}}{\bar{\eta}} \left(\frac{1}{\lambda_{ij}^B} + \frac{1}{t_{ij}} \right)} = 1 - \frac{\lambda_{ij}}{\bar{\eta}} \left(\frac{1}{\lambda_{ij}^B} + \frac{1}{t_{ij}} \right) \equiv 1 - \frac{\lambda_{ij}}{\bar{\eta}} \left(\frac{1}{g(b_{ij})} + \frac{1}{t_{ij}} \right), \quad (\text{A.1})$$

where the first equality uses Equation (7) in Avi-Itzhak and Naor (1963) and the second, the steady-

state flow balance condition that the average rate of inflow to the passenger queue must equal the average rate of outflow, $\lambda_{ij} = P_{ij}^L \tilde{\lambda}_{ij}$. The right-hand side outflow, here, equals the fraction of time that a bus is loading times the expected outflow, conditional on a bus being present.

Finally, I derive queueing and loading times. My queueing time, Q_{ij} , is equivalent to the expected wait time in the queue plus service time in their model, which Avi-Itzhak and Naor (1963) denote by θ_q , and their expected number of customers in the system, q , is equivalent to queue length, n_{ij} , in my model. Avi-Itzhak and Naor (1963) then provide an expression for the expected queue length, $q \equiv n_{ij}$, in Equation (33). Through the lens of my model, this equation applies whenever the hypoexponential distribution of the sum of the time until the bus arrival shock (at rate λ_{ij}^B) and the time until the bus loading start shock (at rate ι_{ij}) has a coefficient of variation, $\gamma \equiv \sqrt{\frac{(\lambda_{ij}^B)^{-2} + \iota_{ij}^{-2}}{(\lambda_{ij}^B)^{-1} + \iota_{ij}^{-1}}}$, equal to 1. For given finite λ_{ij}^B , this condition holds provided interruptions to bus loading are brief, i.e. $\iota_{ij} \rightarrow \infty$, a fair approximation of the relatively fluid loading process. Thus, I substitute their Equation (33) into $Q_{ij} = n_{ij}/\lambda_{ij}$, the equivalent of their Equation (22). Note that their fraction of time the service station is out of service, p_1 , satisfies $p_1 = 1 - p_0 \equiv 1 - P_{ij}^L$, and that (A.1) implies that $\frac{1}{\lambda_{ij}^B} + \frac{1}{\iota_{ij}} = (1 - P_{ij}^L) \frac{\bar{\eta}}{\lambda_{ij}}$. These equivalencies then yield an expression for the expected queueing time, θ_q in their notation, and, in my model,

$$Q_{ij} = \frac{\mu^{-1} + \frac{\bar{\eta}P_{ij}^L(1-P_{ij}^L)^2}{\lambda_{ij}}}{P_{ij}^L - \frac{\lambda_{ij}}{\mu}} \approx \frac{1}{\mu P_{ij}^L - \lambda_{ij}}, \quad (\text{A.2})$$

where the approximation holds for routes with a bus loading most of the time, $P_{ij}^L \rightarrow 1$ such that $(1 - P_{ij}^L)^2 \rightarrow 0$. The bus loading time, in turn, by the properties of the Poisson process, equals the inverse of the rate at which loading buses depart, such that $L_{ij} = \bar{\eta}/\tilde{\lambda}_{ij}$, and, using the aforementioned steady-state relation $\lambda_{ij} = P_{ij}^L \tilde{\lambda}_{ij}$, satisfies

$$L_{ij} = \frac{\bar{\eta}P_{ij}^L}{\lambda_{ij}}. \quad (\text{A.3})$$

A.2 Derivation of Market Mohring Effect

Lemma A.1. *The equilibrium derivative of queueing plus loading time with respect to the passenger arrival rate, λ_{ij} , approximately satisfies*

$$\frac{\partial (Q_{ij} + L_{ij})}{\partial \log \lambda_{ij}} \approx \lambda_{ij} Q_{ij}^2 - L_{ij}$$

for high loading probabilities, $P_{ij}^L \rightarrow 1$.

Proof. Using the chain rule and, as in the main text, applying Assumption 1, the derivative of queueing plus loading time with respect to log passenger arrivals is given by

$$\begin{aligned} \frac{\partial (Q_{ij} + L_{ij})}{\partial \log \lambda_{ij}} &= \left(\frac{\partial \log Q_{ij}}{\partial Q_{ij}} \right)^{-1} \left(\frac{\partial \log Q_{ij}}{\partial \log \lambda_{ij}} + \frac{\partial \log Q_{ij}}{\partial \log P_{ij}^L} \frac{\partial \log P_{ij}^L}{\partial \log \lambda_{ij}} \right) + \\ &\quad \left(\frac{\partial \log L_{ij}}{\partial L_{ij}} \right)^{-1} \left(\frac{\partial \log L_{ij}}{\partial \log \lambda_{ij}} + \frac{\partial \log L_{ij}}{\partial \log P_{ij}^L} \frac{\partial \log P_{ij}^L}{\partial \log \lambda_{ij}} \right) \end{aligned}$$

where both terms reflect partial derivatives with respect to lambda, holding bus supply fixed. From (A.1) and using $\log \left[1 - \frac{\lambda_{ij}}{\bar{\eta}g(b_{ij})} \right] \approx -\frac{\lambda_{ij}}{\bar{\eta}g(b_{ij})}$ for small $\frac{\lambda_{ij}}{\bar{\eta}g(b_{ij})}$, I calculate $\frac{\partial \log P_{ij}^L}{\partial \log \lambda_{ij}} \approx 1 - P_{ij}^L \approx 0$ for $P_{ij}^L \rightarrow 1$. Thus, substituting in for the first derivative in each term, I obtain

$$\frac{\partial (Q_{ij} + L_{ij})}{\partial \log \lambda_{ij}} = Q_{ij} \frac{\partial \log Q_{ij}}{\partial \log \lambda_{ij}} + L_{ij} \frac{\partial \log L_{ij}}{\partial \log \lambda_{ij}}$$

Log-linearizing the approximated version of (A.2), I obtain $\frac{\partial \log Q_{ij}}{\partial \log \lambda_{ij}} \approx \frac{\lambda_{ij}}{\mu P_{ij}^L - \lambda_{ij}} \equiv \lambda_{ij} Q_{ij}$; from (A.3), it is immediate that $\frac{\partial \log L_{ij}}{\partial \log \lambda_{ij}} = -1$. \square

Lemma A.2. *The equilibrium derivative of queueing plus loading time with respect to bus supply, b_{ij} , approximately satisfies*

$$\frac{\partial (Q_{ij} + L_{ij})}{\partial \log b_{ij}} \approx (1 - P_{ij}^L) \varepsilon_{g,b} [-\mu Q_{ij}^2 P_{ij}^L + L_{ij}] .$$

Proof. Once again using the chain rule and Assumption 1, the derivative of queueing plus loading time with respect to log bus supply is given by

$$\begin{aligned} \frac{\partial (Q_{ij} + L_{ij})}{\partial \log b_{ij}} &= \left(\frac{\partial \log Q_{ij}}{\partial Q_{ij}} \right)^{-1} \frac{\partial \log Q_{ij}}{\partial \log b_{ij}} + \left(\frac{\partial \log L_{ij}}{\partial L_{ij}} \right)^{-1} \frac{\partial \log L_{ij}}{\partial \log b_{ij}} \\ &= Q_{ij} \frac{\partial \log Q_{ij}}{\partial \log P_{ij}^L} \frac{\partial \log g(b_{ij})}{\partial \log b_{ij}} + L_{ij} \frac{\partial \log L_{ij}}{\partial \log P_{ij}^L} \frac{\partial \log g(b_{ij})}{\partial \log b_{ij}} \end{aligned}$$

Log-linearizing the approximated version of (A.2), I obtain $\frac{\partial \log Q_{ij}}{\partial \log P_{ij}^L} \approx -\mu Q_{ij} P_{ij}^L$. From (1) and using $\log \left[1 - \frac{\lambda_{ij}}{\bar{\eta}g(b_{ij})} \right] \approx -\frac{\lambda_{ij}}{\bar{\eta}g(b_{ij})}$ for small $\frac{\lambda_{ij}}{\bar{\eta}g(b_{ij})}$, I obtain $\frac{\partial \log P_{ij}^L}{\partial \log g(b_{ij})} \approx 1 - P_{ij}^L$, and I define the

elasticity $\epsilon_{g,b} \equiv \frac{\partial \log g(b_{ij})}{\partial \log b_{ij}}$. Finally, from (A.3), it is immediate that $\frac{\partial \log L_{ij}}{\partial \log P_{ij}^L} = 1$. \square

A.3 Proof of Proposition 1 (Efficiency)

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

$$\max_{\{\pi\}_{i,j,m,g} \{b\}_{i,j}} \Omega \text{ subject to (1)-(3), (9), and } \sum_{i,j,m} \pi_{ijm}^g = 1. \quad (\text{A.4})$$

I substitute (1)-(3) and (9) into the expression in Lemma C.1 in Online Appendix C.1 and apply Assumption 1 to obtain welfare as $\Omega(b, \pi)$. I can then rewrite the planner's problem as $\max_{b, \pi} \Omega(b, \pi)$ s.t. $\sum_{i,j,m} \pi_{ijm}^g = 1$.²⁶ The planner's first-order conditions for the mass of searching buses b_{ij} on each route read

$$\frac{\partial \Omega}{\partial b_{ij}} = \sum_g N^g \pi_{ijM}^{g*} \left\{ -r \omega_j^g \left[\frac{\partial Q_{ij}}{\partial b_{ij}} + \frac{\partial L_{ij}}{\partial b_{ij}} \right] - \frac{\bar{\omega}}{\bar{\eta}} \frac{\partial L_{ij}}{\partial b_{ij}} \right\} - \bar{\omega} = 0. \quad (\text{A.5})$$

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_{klA}^{g*} , to derive an optimal relative choice probability:

$$\begin{aligned} \log \left(\pi_{ijm}^{g*} / \pi_{klA}^{g*} \right) &= \exp \left(\theta_i^g - \kappa_m^g + \omega_j^g \right. \\ &\quad \left. + 1 \{m = M\} \left\{ -r \omega_j^g (Q_{ij} + L_{ij} + T_{ij}) - \frac{\chi_{ij}}{\bar{\eta}} - \frac{\bar{\omega}}{\bar{\eta}} (L_{ij} + T_{ij}) \right. \right. \\ &\quad \left. \left. - \sum_{g'} N^{g'} \pi_{ijM}^{g'*} \left[r \omega_j^{g'} \left(\frac{\partial Q_{ij}}{\partial \lambda_{ij}} + \frac{\partial L_{ij}}{\partial \lambda_{ij}} \right) + \frac{\bar{\omega}}{\bar{\eta}} \frac{\partial L_{ij}}{\partial \lambda_{ij}} \right] \right\} \right. \\ &\quad \left. + 1 \{m = F\} \left[-r \omega_j^g (H_{ij} + T_{ijF}) - \tau_{ijF} \right] + 1 \{m = A\} \left(-r \omega_j^g T_{ij} - \tau_A \right) \right. \\ &\quad \left. - \theta_k^g + \kappa_A^g + r \omega_l^g T_{kl} + \tau_A - \omega_l^g \right)^{1/\nu}. \quad (\text{A.6}) \end{aligned}$$

Equations (A.5), (A.6), and the adding-up constraints in (A.4) 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_{ijm}^{g*} .

²⁶This welfare function includes the expectation of the idiosyncratic shocks, under the assumption that these choice probabilities can be implemented through appropriate transfers to commuters. I later confirm that the social planner can indeed do so.

Second, I derive conditions under which the social planner solution defined by (A.5)-(A.6) coincides with the equilibrium defined in the main text. Starting with bus supply, I substitute profits (4) into the Nash-bargaining equation (6) and then into the bus supply first-order condition (5). The resulting equation,

$$\underbrace{\left\{ \frac{\omega_j + \frac{\bar{\omega}(1-\beta)\lambda_{ij}}{\beta \frac{\partial \Pi_{ij}}{\partial \tau_{ij}}} \left(\frac{\chi_{ij}}{\eta \bar{\omega}} + \frac{L_{ij}+T_{ij}}{\eta} + \frac{b_{ij}}{\lambda_{ij}} \right)}{1 + \frac{(1-\beta)\lambda_{ij}}{\beta \frac{\partial \Pi_{ij}}{\partial \tau_{ij}}}} - \frac{\chi_{ij}}{\eta} - \frac{\bar{\omega}}{\eta} (L_{ij} + T_{ij}) \right\}}_{\equiv \tau_{ij}} \sum_g N^g \frac{\partial \pi_{ijM}^g}{\partial b_{ij}} - \frac{\lambda_{ij}}{\eta} \bar{\omega} \left(\frac{\partial L_{ij}}{\partial \lambda_{ij}} \sum_g N^g \frac{\partial \pi_{ijM}^g}{\partial b_{ij}} + \frac{\partial L_{ij}}{\partial b_{ij}} \right) - \bar{\omega} = 0, \quad (\text{A.7})$$

determines equilibrium association bus supply on each route. This equilibrium supply b_{ij} coincides with the optimal entry b_{ij}^* pinned down by the planner's bus entry condition (A.5) if and only if

$$\underbrace{\left\{ \frac{\omega_j + \frac{\bar{\omega}(1-\beta)\lambda_{ij}}{\beta \frac{\partial \Pi_{ij}}{\partial \tau_{ij}}} \left(\frac{\chi_{ij}}{\eta \bar{\omega}} + \frac{L_{ij}+T_{ij}}{\eta} + \frac{b_{ij}}{\lambda_{ij}} \right)}{1 + \frac{(1-\beta)\lambda_{ij}}{\beta \frac{\partial \Pi_{ij}}{\partial \tau_{ij}}}} - \frac{\chi_{ij}}{\eta} - \frac{\bar{\omega}}{\eta} (L_{ij} + T_{ij}) \right\}}_{\equiv \tau_{ij}} \sum_g N^g \frac{\partial \pi_{ijM}^g}{\partial b_{ij}} - \frac{\lambda_{ij}}{\eta} \bar{\omega} \frac{\partial L_{ij}}{\partial \lambda_{ij}} \sum_g N^g \frac{\partial \pi_{ijM}^g}{\partial b_{ij}} = - \sum_g N^g \pi_{ijM}^g r \omega_j^g \left[\frac{\partial Q_{ij}}{\partial b_{ij}} + \frac{\partial L_{ij}}{\partial b_{ij}} \right]. \quad (\text{A.8})$$

Next, turning to demand, I substitute commute values U_{ijm}^g into (8); it is then immediate that equilibrium (relative) choice probabilities, $\log(\pi_{ijm}^g / \pi_{kla}^g)$, equal those chosen by the planner in (A.6) if and only if bargained minibus fares compensate, for each route and skill group, per-passenger operations costs and driver wages plus the marginal value of the Direct Market Mohring Effect:

$$\underbrace{\left\{ \frac{\omega_j + \frac{\bar{\omega}(1-\beta)\lambda_{ij}}{\beta \frac{\partial \Pi_{ij}}{\partial \tau_{ij}}} \left(\frac{\chi_{ij}}{\eta \bar{\omega}} + \frac{L_{ij}+T_{ij}}{\eta} + \frac{b_{ij}}{\lambda_{ij}} \right)}{1 + \frac{(1-\beta)\lambda_{ij}}{\beta \frac{\partial \Pi_{ij}}{\partial \tau_{ij}}}} = \frac{\chi_{ij}}{\eta} + \frac{\bar{\omega}}{\eta} (L_{ij} + T_{ij}) \right\}}_{\equiv \tau_{ij}} + \sum_{g'} N^{g'} \pi_{ijM}^{g'} \left[r \omega_j^{g'} \left(\frac{\partial Q_{ij}}{\partial \lambda_{ij}} + \frac{\partial L_{ij}}{\partial \lambda_{ij}} \right) + \frac{\bar{\omega}}{\eta} \frac{\partial L_{ij}}{\partial \lambda_{ij}} \right]. \quad (\text{A.9})$$

□

A.4 Derivation of Queueing Efficiency Estimation Equation

I now derive Equation (15), used in the estimation of the queueing efficiency μ . Note that, for purposes of estimation, I do not apply Assumption 1 and instead allow for bus loading interruptions to arrive at rate $\iota_{ij} < \infty$. Avi-Itzhak and Naor (1963) discuss the well-known result known as Little's Law that the average number of customers (i.e. passengers) waiting in a queue satisfies, in my notation, $n_{ij} = \lambda_{ij}Q_{ij}$. I combine this insight with (A.2) and divide numerator and denominator by $\bar{\eta}$ to find that

$$n_{ij} = \frac{\lambda_{ij}(\bar{\eta}\mu)^{-1} + P_{ij}^L \left(1 - P_{ij}^L\right)^2}{\frac{P_{ij}^L}{\bar{\eta}} - \frac{\lambda_{ij}}{\mu\bar{\eta}}} \approx \frac{\lambda_{ij}(\bar{\eta}\mu)^{-1}}{\frac{P_{ij}^L}{\bar{\eta}} - \frac{\lambda_{ij}}{\mu\bar{\eta}}} \quad (\text{A.10})$$

where the approximation again follows from the fact that $\left(1 - P_{ij}^L\right)^2 \approx 0$ for P_{ij}^L close to 1. I then multiply numerator and denominator by μ/λ_{ij} and take logs to obtain

$$\log n_{ij} \approx \log\left(\frac{1}{\bar{\eta}}\right) - \log\left(\frac{\mu P_{ij}^L}{\lambda_{ij}\bar{\eta}} - \frac{1}{\bar{\eta}}\right) \approx -\mu \underbrace{\frac{P_{ij}^L}{\lambda_{ij}\bar{\eta}}}_{\equiv h(\bar{\eta})} + \underbrace{\log\left(\frac{1}{\bar{\eta}}\right) + \frac{1}{\bar{\eta}} + 1}_{\equiv h(\bar{\eta})}, \quad (\text{A.11})$$

where the second approximation uses the fact that $\frac{\mu P_{ij}^L}{\lambda_{ij}\bar{\eta}} - \frac{1}{\bar{\eta}} - 1$ is close to zero given median values of each variable in my data and reasonable values of μ . In consequence, $\log\left(\frac{\mu P_{ij}^L}{\lambda_{ij}\bar{\eta}} - \frac{1}{\bar{\eta}}\right) \approx \frac{\mu P_{ij}^L}{\lambda_{ij}\bar{\eta}} - \frac{1}{\bar{\eta}} - 1$.

However, I do not observe the disruptions, captured by $\iota_{ij} < \infty$, to the loading process that would permit me to compute P_{ij}^L and estimate (A.11) as written. Instead, from the rank counts, I observe the probability that a bus is present in the loading lane, which I denote by $\Pr(\text{bus dwell})_{ij}$, where this probability includes both the case where a bus is present but has not yet received the loading shock and when a bus is actively loading. Thus, by the properties of Poisson processes,

$$\Pr(\text{bus dwell})_{ij} = \frac{\frac{1}{\iota_{ij}} + L_{ij}}{\frac{1}{\lambda_{ij}^B} + \frac{1}{\iota_{ij}} + L_{ij}} = P_{ij}^L + \frac{\lambda_{ij}}{\bar{\eta}\iota_{ij}}, \quad (\text{A.12})$$

where the second equality uses the fact that (A.1) implies that $\frac{1}{\lambda_{ij}^B} + \frac{1}{\iota_{ij}} = \left(1 - P_{ij}^L\right) \frac{\bar{\eta}}{\lambda_{ij}}$ as well as (A.1) and (A.3). I then substitute the expression for P_{ij}^L implied by (A.12) into (A.11) to obtain the

estimating Equation (15) in the main text,

$$\log n_{ij} \approx -\mu \frac{\Pr(\text{bus dwell})_{ij}}{\lambda_{ij}\bar{\eta}} + h(\bar{\eta}) + \underbrace{\frac{\mu}{\lambda_{ij}\bar{\eta}^2}}_{\equiv \varepsilon_{ij}}, \quad (\text{A.13})$$

where the composite error term ε_{ij} captures the unobserved disruptions to bus loading.

A.5 Alternative Market Structure: Free Entry of Minibuses

In this section, I lay out the free entry market structure, which underlies my third policy counterfactual. First, I assume that individual minibuses Nash bargain fares jointly with their full busload of $\bar{\eta}$ passengers after the last passenger boards and before departure; I denote bus bargaining power by $\tilde{\beta}$. If negotiations fail, passengers go to the back of the queue and resume waiting. Since, in stationary equilibrium, passengers receive exactly the same value once they return to the front of the queue, each passenger's expected surplus from bargaining simply equals the value of (expected) time saved by not cycling through the queue again, $r\omega_j^g(Q_{ij} + L_{ij}) - (\tau^{\text{FE}} - \tau_{ij}^{\text{oo}})$, net of the excess of the bargained fare τ_{ij}^{FE} over the fare passengers would pay upon returning to the front of the queue, τ_{ij}^{oo} . As for buses, I assume that they must exit the market and receive zero surplus if they fail to reach an agreement, so their surplus per trip equals the revenue minus operating costs and driver opportunity cost of travel time.

Thus, the free-entry-model fares solve

$$\max_{\tau_{ij}^{\text{FE}}} (\bar{\eta} \tau_{ij}^{\text{FE}} - \chi_{ij} - \bar{\omega} T_{ij})^{\tilde{\beta}} \left[\bar{\eta} r E(\omega_j^g) (Q_{ij} + L_{ij}) - (\tau^{\text{FE}} - \tau_{ij}^{\text{oo}}) \right]^{1-\tilde{\beta}} \quad (\text{A.14})$$

where I have assumed that passengers bargain based on their average wage, $E(\omega_j^g) \equiv \frac{\sum_g N^g \pi_{ijM}^g \omega_j^g}{\sum_g N^g \pi_{ijM}^g}$. Note that queueing and loading times do not depend on the fare bargained by an individual bus. Taking first-order condition and imposing the equilibrium condition $\tau^{\text{FE}} = \tau_{ij}^{\text{oo}}$, I obtain the free-entry fare as

$$\tau_{ij}^{\text{FE}} = \frac{\chi_{ij}}{\bar{\eta}} + \frac{\bar{\omega}}{\bar{\eta}} T_{ij} + \frac{\tilde{\beta}}{1-\tilde{\beta}} r E(\omega_j^g) (Q_{ij} + L_{ij}). \quad (\text{A.15})$$

Profits per bus, using the free-entry fares and (4), then satisfy

$$\frac{\Pi_{ij}}{b_{ij}} = \frac{\lambda_{ij}}{b_{ij}\bar{\eta}} \left[\frac{\tilde{\beta}}{1-\tilde{\beta}} \bar{\eta} r E(\omega_j^g) (Q_{ij} + L_{ij}) - \bar{\omega} L_{ij} \right] - \bar{\omega} = 0 \quad (\text{A.16})$$

where the final zero-profits equality results exactly from free entry of minibuses.

The remaining equilibrium equations are as before, so that the equilibrium definition is as in the main text, except that fares τ_{ij}^{FE} and bus supply b_{ij} are determined by (A.15) and (A.16) instead of (6) and (5).

REFERENCES

- 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.
- Akbar, Prottoy A., Victor Couture, Gilles Duranton, and Adam Storeygard. 2023a. “Mobility and Congestion in Urban India.” *American Economic Review* 113 (4): 1083–1111.
- . 2023b. “The fast, the slow, and the congested: Urban transportation in rich and poor countries.” *Revise & Resubmit, Quarterly Journal of Economics*.
- Allen, Treb, and Costas Arkolakis. 2014. “Trade and the Topography of the Spatial Economy.” *The Quarterly Journal of Economics* 129 (3): 1085–1140.
- . 2022. “The Welfare Effects of Transportation Infrastructure Improvements.” *The Review of Economic Studies* 89 (6): 2911–2957.
- Almagro, Milena, Felipe Barbieri, Juan Camilo Castillo, Nathaniel Hickok, and Tobias Salz. 2024. “Optimal Urban Transportation Policy: Evidence from Chicago.”
- Ameriks, John, Joseph Briggs, Andrew Caplin, Matthew D. Shapiro, and Christopher Tonetti. 2020. “Long-Term-Care Utility and Late-in-Life Saving.” *Journal of Political Economy* 128 (6): 2375–2451.
- Andrew, Alison, and Abi Adams-Prassl. 2023. “Revealed Beliefs and the Marriage Market Return to Education.”
- Andrews, Isaiah, James Stock, and Liyang Sun. 2019. “Weak Instruments in IV Regression: Theory and Practice.” *Annual Review of Economics* 11:727–753.
- Antrobus, Lauren, and Andrew Kerr. 2019. “The labour market for minibus taxi drivers in South Africa.” SALDRU Working Paper No. 250.
- Asanjarani, Azam, Yoni Nazarathy, and Peter Taylor. 2021. “A survey of parameter and state estimation in queues.” *Queueing Systems* 97:39–80.
- 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.”
- 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.
- Borck, Rainald. 2019. “Public transport and urban pollution.” *Regional Science and Urban Economics* 77:356–366.
- 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. 2020a. "A guide to estimating matching functions in spatial models." *International Journal of Industrial Organization* 70:102533.
- . 2020b. "Geography, Transportation, and Endogenous Trade Costs." *Econometrica* 88 (2): 657–691.
- . 2024. "Investment in Infrastructure and Trade: The Case of Ports."
- Brancaccio, Giulia, Myrto Kalouptsidi, Theodore Papageorgiou, and Nicola Rosaia. 2023. "Search Frictions and Efficiency in Decentralized Transport Markets." *Forthcoming, Quarterly Journal of Economics*.
- 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*.
- Castillo, Juan Camilo. 2022. "Who Benefits from Surge Pricing?" *Conditionally accepted, Econometrica*.
- Cervero, Robert, and Aaron Golub. 2007. "Informal transport: A global perspective." *Transport policy* 14 (6): 445–457.
- Chowdhury, Shovan, and S.P. Mukherjee. 2011. "Estimation of waiting time distribution in an M/M/1 Queue." *Opsearch* 48 (4): 306–317.
- . 2013. "Estimation of Traffic Intensity Based on Queue Length in a Single M/M/1 Queue." *Communications in Statistics—Theory and Methods* 42:2376–2390.
- 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 Krogsoepers, 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.
- Donaldson, Dave. 2018. "Railroads of the Raj: Estimating the Impact of Transportation Infrastructure." *American Economic Review* 108 (4–5): 899–934.
- Donaldson, Dave, and Richard Hornbeck. 2016. "Railroads and American Economic Growth: A "Market Access" Approach." *The Quarterly Journal of Economics* 131 (2): 799–858.
- Duranton, Gilles, and Matthew A. Turner. 2011. "The Fundamental Law of Road Congestion: Evidence from US Cities." *American Economic Review* 101 (6): 2616–2652.
- Fajgelbaum, Pablo D., Amit Khandelwal, Wookun Kim, Cristiano Mantovani, and Edouard Schaal. 2021. "Optimal Lockdown in a Commuting Network." *American Economic Review: Insights* 3 (4): 503–522.
- Fajgelbaum, Pablo D., and Edouard Schaal. 2020. "Optimal Transport Networks in Spatial Equilibrium." *Econometrica* 88 (4): 1411–1452.
- Fuchs, Simon, and Woan Foong Wong. 2024. "Multimodal Transport Networks."
- Glaeser, Edward. 2020. "Infrastructure and Urban Form." NBER Working Paper 28287.

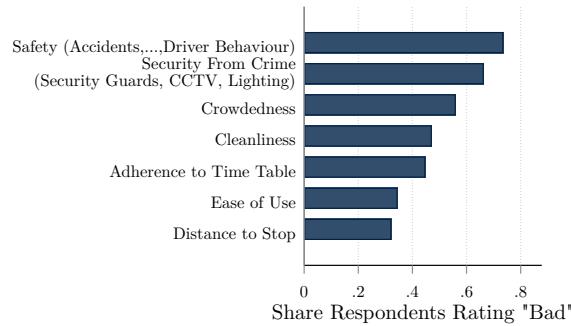
- Goldszmidt, Ariel, John A. List, Robert D. Metcalfe, Ian Muir, V. Kerry Smith, and Jenny Wang. 2020. "The Value of Time in the United States: Estimates from Nationwide Natural Field Experiments."
- Heblich, Stephan, Stephen J. Redding, and Daniel M. Sturm. 2020. "The Making of the Modern Metropolis: Evidence from London." *The Quarterly Journal of Economics*, 2059–2133.
- 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.
- Jobanputra, Rahul. 2018. *Comprehensive Integrated Transport Plan 2018 – 2023*. Technical report. City of Cape Town Transport and Urban Development Authority, January. <https://tdacontenthub.bstorer.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.
- 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.
- Meyer, J. R. ad J.F. Kain, and M. Wohl. 1965. *The Urban Transportation Problem*. Cambridge, MA and London, England: Harvard University Press.
- Mohring, Herbert. 1972. "Optimization and Scale Economies in Urban Bus Transportation." *American Economic Review* 62 (4): 591–604.
- . 1983. "Minibuses in urban transportation." *Journal of Urban Economics* 14 (3): 293–317.
- Monte, Ferdinando, Stephen J. Redding, and Esteban Rossi-Hansberg. 2018. "Commuting, Migration, and Local Employment Elasticities." *American Economic Review* 108 (12): 3855–3890.
- Nagy, Dávid Krisztián. 2023. "Hinterlands, city formation and growth: Evidence from the U.S. westward expansion." *Forthcoming, Review of Economic Studies*.
- Oldfield, R.H. ad P.H. Bly. 1988. "An analytic investigation of optimal bus size." *Transportation Research Part B: Methodological* 22 (5): 319–337.
- 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, Herrie. 2017. "Lessons from building paratransit operators' capacity to be partners in Cape Town's public transport reform process." *Transportation Research Part A* 104:58–66.
- Severen, Christopher. 2023. "Commuting, Labor, and Housing Market Effects of Mass Transportation: Welfare and Identification." *Forthcoming, Review of Economics and Statistics*.
- Small, Kenneth A., and Erik T. Verhoef. 2007. *The Economics of Urban Transportation*. New York: Routledge.
- Theway, Chesway. 2018. *Pros & Cons of Minibus Taxis: The Transport System in South Africa*. <https://theway.medium.com/pros-cons-of-minibus-taxis-23af16de783>.
- 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>.
- U.S. Department of Energy. 2023. *Alternative Fuels Data Center: Public Transportation*. Technical report. https://afdc.energy.gov/conserve/public_transportation.html.
- Walters, A.A. 1979. "The Benefits of Minibuses: The Case of Kuala Lumpur." *Journal of Transport Economics and Policy* 13 (3): 320–334.
- Warnes, Pablo Ernesto. 2021. "Transport Infrastructure Improvements and Spatial Sorting: Evidence from Buenos Aires."
- White, P.R. ad R.P. Turner. 1987. "Development Of Intensive Urban Minibus Services In Britain." *Logistics ad Transportation Review* 23 (4): 385–400.
- 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

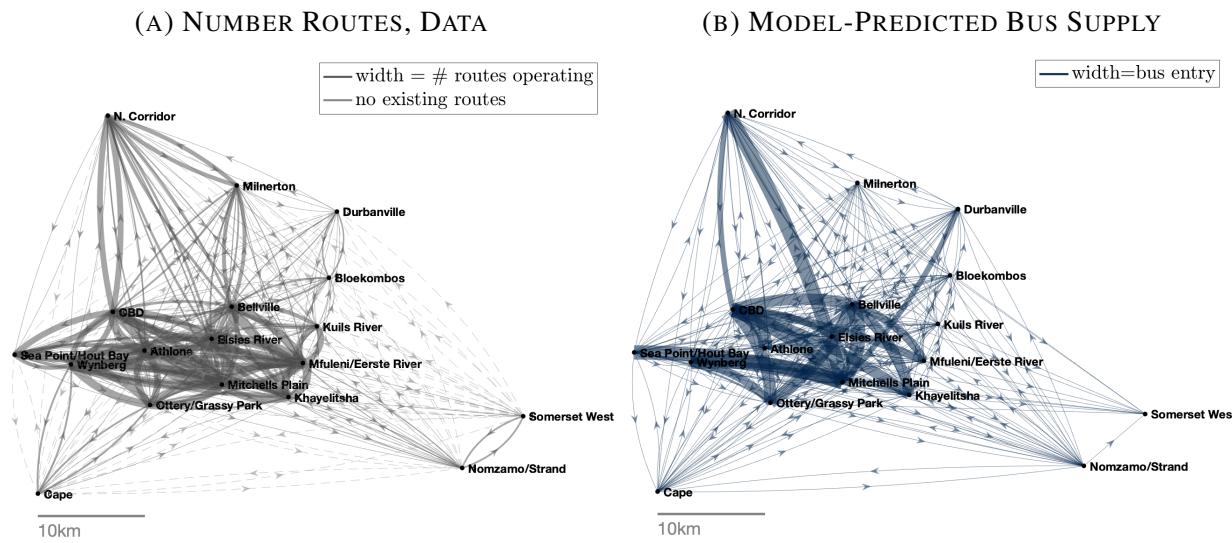
A. ADDITIONAL FIGURES

FIGURE A.1. RATINGS OF MINIBUS ATTRIBUTES



Notes: Figure displays 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.”

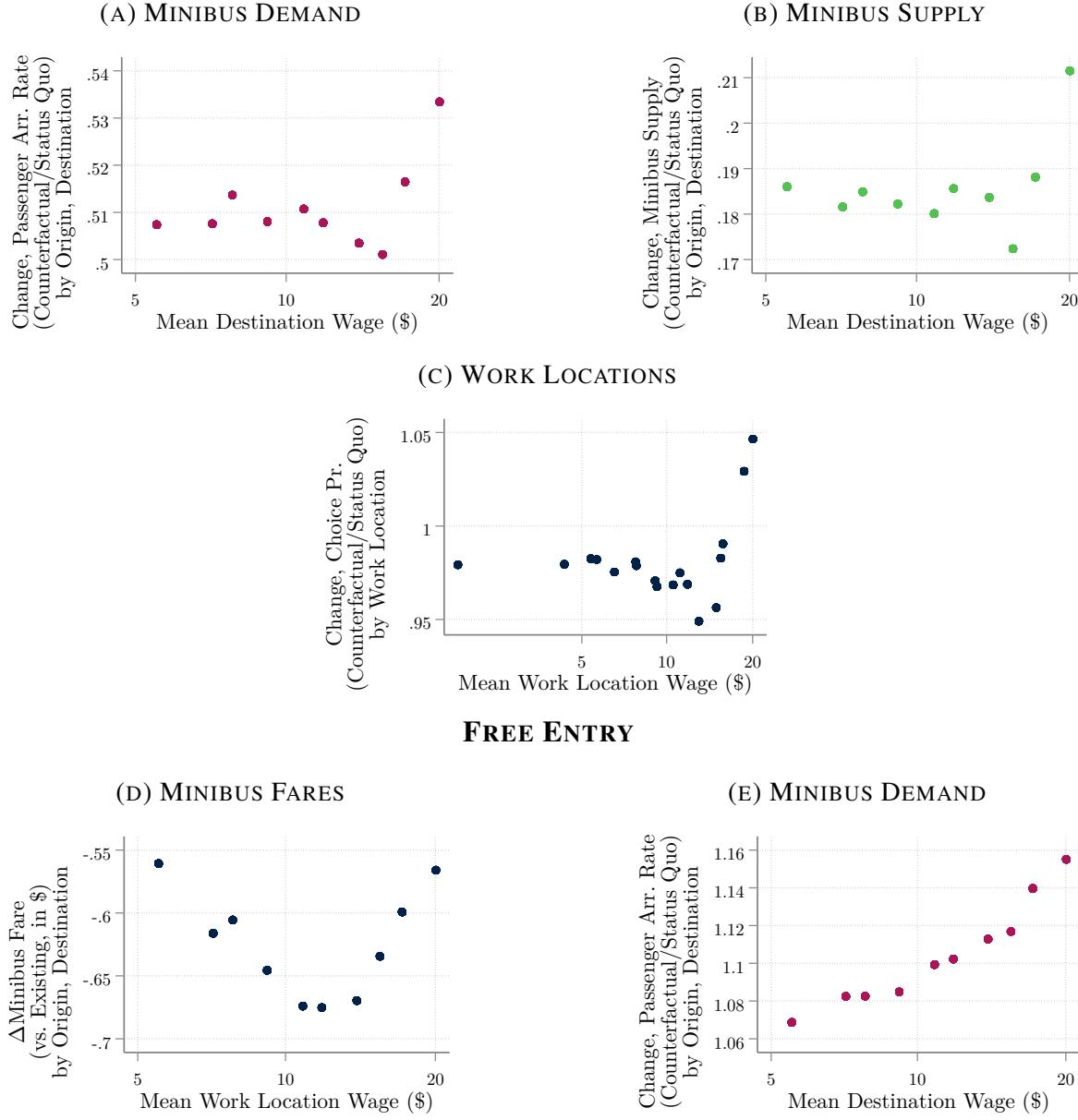
FIGURE A.2. MINIBUS NETWORK IN DATA VERSUS MODEL



Notes: Map in Panel (A) displays 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.

FIGURE A.3. CHANGES IN MARKET STRUCTURE

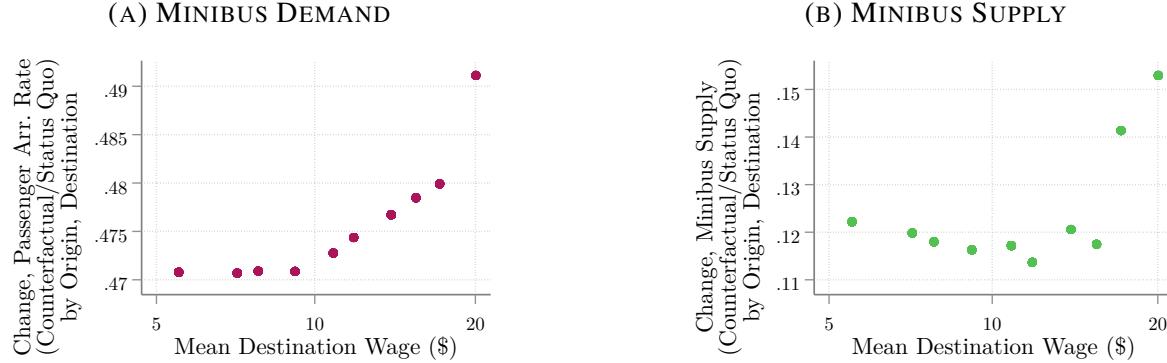
MONOPOLY PRICING



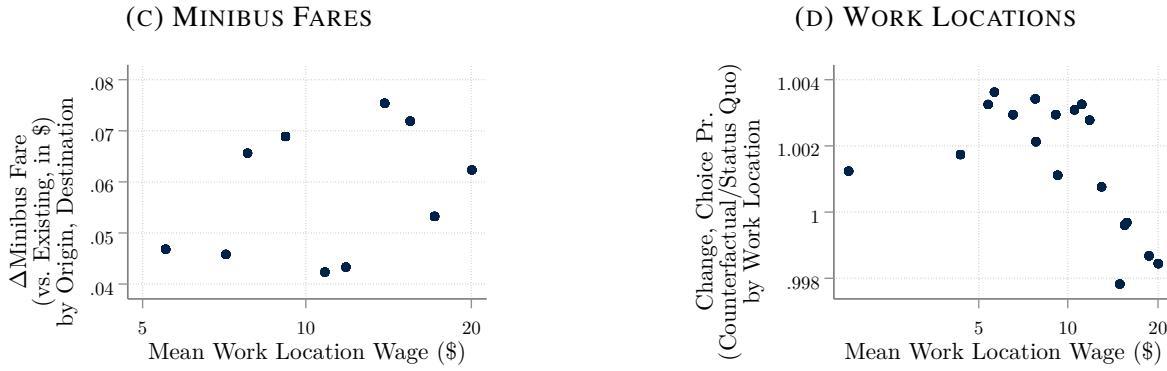
Notes: Panels (A) and (E) display binned scatterplots of the commuter inflow λ_{ij} under a given market structure counterfactual, in changes relative to the status quo at the route level, on the vertical axis. The horizontal axis displays the average across skill groups, weighted by aggregate populations, of the corresponding work-location wage. Panel (B) instead displays the changes in minibus supply b_{ij} . Panel (C) displays a scatterplot of the proportionate change in work location choice probabilities in a given counterfactual, averaged over skill groups, versus that work location's average wage. Panel (D) displays another binned scatterplot of the raw changes in route-level minibus fares.

FIGURE A.4. CHANGES IN TECHNOLOGY

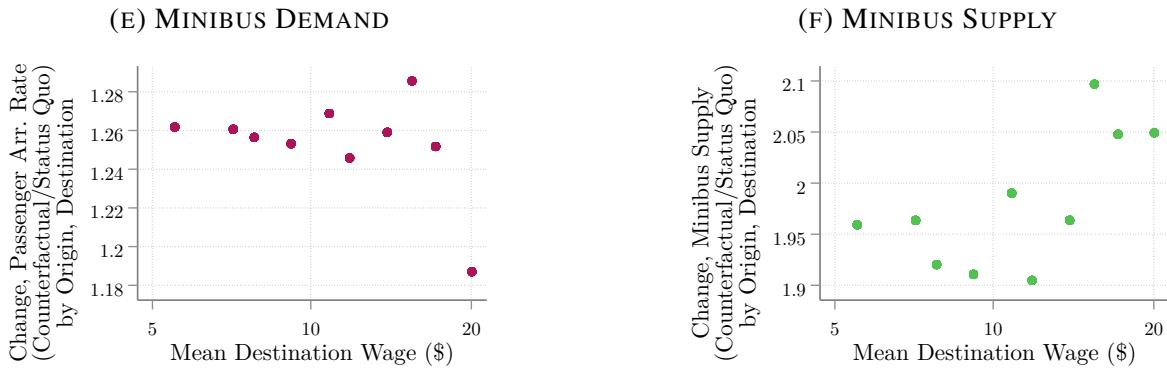
TWO QUEUES



SMALLER MINIBUSES



MINIBUS SECURITY



Notes: Panels (A) and (E) display binned scatterplots of the commuter inflow λ_{ij} under a given market structure counterfactual, in changes relative to the status quo at the route level, on the vertical axis. The horizontal axis displays the average across skill groups, weighted by aggregate populations, of the corresponding work-location wage. Panels (B) and (F) instead display the changes in minibus supply b_{ij} , and Panel (C), the raw changes in route-level minibus fares. Panel (D) displays a scatterplot of the proportionate change in work location choice probabilities in a given counterfactual, averaged over skill groups, versus that work location's average wage.

B. DATA

B.1 Minibus Station Counts

Here, I provide more detail on the minibus station counts used to characterize the matching process. I designed the counts in cooperation with input from the mobility advisory firm GoAscendal, who 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.²⁷ 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 519 routes operating from 107 stations. Of these stations, 31 have at least 6 routes, and these stations account for 328 of the 519 total routes, leaving a total population of $N = 328$ routes.

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 entry, namely the aforementioned number of trips mapped in the previous data collection effort, to over-sample stations and routes with higher levels of bus entry and thus reduce the number of zero bus and passenger observations in the resulting data.

²⁷In 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 (Coetze et al. (2018)).

In the first stage, I took a stratified random sample of stations. First, I took a 100% sample of the 5 highest-bus-traffic feasible stations that operate in the morning peak, as measured by my proxy of bus entry, i.e. the total trips mapped originating at that station in the previous 2018 study.²⁸ Second, I sampled 3 stations, or a 16% sample, from the remaining 19 non-duplicate lower-bus-traffic stations with no permission issues, 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.²⁹

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 entry, 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.

Practical Implementation 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 there were additional routes 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.

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

²⁸From 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 in fact numbers 1-2, 4, 6, and 8.

²⁹Wynberg 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.

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
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.

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.³⁰ 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 one or 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 . Denote by $dwelltime_{sl}$ the number of minutes between the time the minibus for trip s on route l arrived to the front of the loading bay and the time it departs, as described in Online Appendix Section B.1.2. I assume that the passengers I observe departing from the origin station on trip s , $deppax_{sl}$, board the bus at a uniform rate. Then, I proportionately apportion these passengers

³⁰ 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.

who depart on a trip to the five-minute blocks during which the bus was dwelling at the front of the loading bay to calculate the total number of passengers boarding buses on route l from t to $t + 5$ as

$$\text{loading passengers}_{lt} \equiv \sum_s \frac{\text{dwelltime}_{sl} \cap [t, t + 5]}{\text{dwelltime}_{sl}} \text{deppax}_{sl}.$$

Here, $\text{dwelltime}_{sl} \cap [t, t + 5]$ indicates the number of minutes of trip s 's dwell time that overlap temporally with clock times t to $t + 5$. I observe the queue length in passengers, n_{lt} , at the beginning of every 5-minute block and then calculate the number of newly arriving passengers per minute as

$$\text{newly-arriving passengers}_{lt} \equiv \frac{n_{l,t+1} + \text{loading passengers}_{lt} - n_{lt}}{5}$$

under the assumption that no passengers abandon the queue after joining. The mean queueing time of these passengers to board buses, consistent with any stationary queue, is $Q_{lt} \equiv 5 \cdot \frac{n_{lt}}{\text{loading passengers}_{lt}}$ and is measured in minutes. I count the number of buses observed arriving to the front of the loading bay within the 5-minute block to obtain the per-minute *newly-arriving buses* _{lt} measure. My dwelltime_{sl} measure technically includes both actual “loading time” when passengers were actively boarding, as defined through the lens of the model, and any disruptions to this loading process. For the facts and model validation, however, I abstract away from disruptions to the loading process – which I do not observe – and calculate loading time L_{lt} as the average observed dwelltime_{sl} across buses departing within the time period.

Second, I calculate the quantities needed to estimate queueing efficiency, as in Equation. (15). At the route ij by hour t level, I calculate averages of (i) the aforementioned queue length to obtain n_{ijt} in (15), (ii) the newly-arriving passengers per minute to estimate λ_{ijt} , and (iii) the average number of passengers onboard a departing bus within that hour to obtain $\bar{\eta}_{ijt}$. I calculate $\Pr(\text{bus dwell})_{ij}$ as the share of minutes during an hour during which I observe a bus present at the front of the loading bay.

Finally, I calculate variables for estimation of the minibus arrivals function and internal calibration. I use the vehicle IDs of buses observed arriving to the station on a given route during a given hour to calculate the number of unique buses b_{ijt} supplied to the route. The route-by-hour average of the minutes elapsed between the departure of one bus and when another bus arrives to the front of the loading bay yields $\overline{\text{bus gap}}_{ijt}$. Using the onboard tracking data, I calculate the median minibus fare for internal calibration by first taking a passenger-weighted average by route and then taking the median across routes.

B.2 Minibus Stated Preference Survey

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

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 cost.

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.”³¹

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.³²

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 of a discrete choice model under some 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

³¹Other aspects included timetable adherence, cleanliness, distance to stop, and ease of use.

³²Fares 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.

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 I now require priors for the demand model parameters. I obtain priors $v = 12.7$, $r = 0.002$, and $\kappa_M = 0.88$ from estimating a mode choice model on the Cape Town household travel survey stated preference module.³³ Then, I use the results of my 1-day stated preference pilot survey (with a very similar format, $N = 20$) to estimate priors for ξ_z , $z \in \{\text{security, no speeding, no overloading}\}$, obtaining $\xi_{\text{security}} = -0.18$, $\xi_{\text{no speeding}} = -0.16$, and $\xi_{\text{no overloading}} = -0.21$.³⁴ I set $\theta_z = -0.1$ 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.³⁵ 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

³³I 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.

³⁴In 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.

³⁵I 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.

late to work!” However, after the pilot survey, I reduced the 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 transport interchange and at 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 formal 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 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.³⁶

B.2.4 Administration and Script

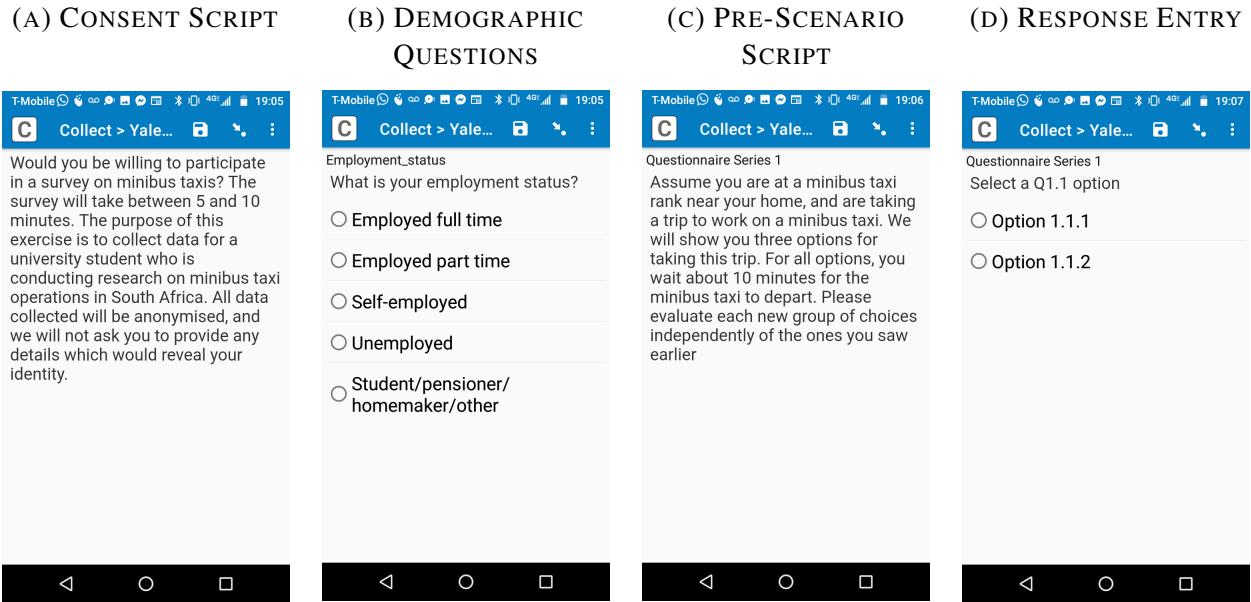
Survey enumerators randomly approached respondents and asked their consent to participate in the survey, according to a script in Online Appendix Figure B.6a. They were offered a chocolate as an incentive. Enumerators then proceeded to read them questions shown in the Survey CTO Android app (see Online 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.

³⁶Enumerators 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



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.

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.1 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. I later pursue multiple strategies to quantify any bias resulting from this oversampling of minibus users.

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

TABLE B.1. STATED PREFERENCE SAMPLE CHARACTERISTICS

Variable	Stated Pref. Samples		Data
	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 [bin]	\$182-\$364	\$182-\$364	\$182-\$364
Median Age	35	39	39
<i>Commute Mode Shares of...</i>			
Minibus	59.56	22.56	23.55
Formal Transit	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.

includes any commuter who uses minibuses during his or her commute; formal transit 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 formal transit; 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: car, formal MyCiti bus, (regular) formal bus, formal train, or minibus (taxi). However, this city-run survey did not include non-pecuniary quality improvements such as station security.

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 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, 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. ADDITIONAL DERIVATIONS AND MICRO-FOUNDATIONS

C.1 Alternative Welfare Expression

Lemma C.1. *Welfare, under Assumption 1, satisfies*

$$\begin{aligned}\Omega = \sum_{i,j,g} N^g \pi_{ijM}^g & \left[\theta_i^g - \kappa_M^g - r\omega_j^g (Q_{ij} + L_{ij} + T_{ij}) + \omega_j^g - v \log \pi_{ijM}^g - \frac{\chi_{ij}}{\bar{\eta}} - \frac{\bar{\omega}}{\bar{\eta}} (L_{ij} + T_{ij}) \right] \\ & + \sum_{i,j,g} N^g \pi_{ijF}^g \left[\theta_i^g - \kappa_F^g - r\omega_j^g (H_{ij} + T_{ijF}) - \tau_{ijF} + \omega_j^g - v \log \pi_{ijF}^g \right] \\ & + \sum_{i,j,g} N^g \pi_{ijA}^g \left[\theta_i^g - \kappa_A^g - r\omega_j^g T_{ij} - \tau_A + \omega_j^g - v \log \pi_{ijA}^g \right] - \bar{\omega} b_{ij}. \quad (\text{C.1})\end{aligned}$$

Proof. In (12), consider first the ex-ante expected utility of group- g commuters, \bar{W}^g , given their optimal choices of home, work, and mode, subject to idiosyncratic Gumbel-distributed preference shocks.³⁷ Denoting total deterministic utility of alternative ijm by $\bar{U}_{ijm}^g \equiv \theta_i^g + U_{ijm}^g + \omega_j^g$, I rewrite expected utility as

$$\begin{aligned}\bar{W}^g & \equiv E \left[\max_{i',j',m'} \left(\bar{U}_{i'j'm'}^g + v \epsilon_{i'j'm'} \right) \right] \\ & = \sum_{i,j,m} \pi_{ijm}^g \left[\bar{U}_{ijm}^g + v E \left(\epsilon_{ijm} | ijm \in \operatorname{argmax}_{i',j',m'} \left(\bar{U}_{i'j'm'}^g + v \epsilon_{i'j'm'} \right) \right) \right] \\ & = \sum_{i,j,m} \pi_{ijm}^g \left[\theta_i^g + U_{ijm}^g + \omega_j^g - v \log \pi_{ijm}^g \right]. \quad (\text{C.2})\end{aligned}$$

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.³⁸ Next, I substitute (9) into association profits (4) and sum across routes to

³⁷Note that the social planner will choose choice probabilities directly and then implement these choice probabilities with appropriately-set transfers.

³⁸To see this, note that

$$\begin{aligned}E \left[\epsilon_{ijm} | ijm \in \operatorname{argmax}_{i',j',m'} \left(\bar{U}_{i'j'm'}^g + v \epsilon_{i'j'm'} \right) \right] & = \frac{1}{v} \left[E \left[\max_{i',j',m'} \left(\bar{U}_{i'j'm'}^g + v \epsilon_{i'j'm'} \right) \right] - \bar{U}_{ijm}^g \right] \\ & = \frac{1}{v} \left[v \log \left[\sum_{i',j',m'} \exp \left[\bar{U}_{i'j'm'}^g \right]^{1/v} \right] - \bar{U}_{ijm}^g \right] = \log \left[\sum_{i',j',m'} \exp \left[\bar{U}_{i'j'm'}^g \right]^{1/v} / \exp \left(\bar{U}_{ijm}^g \right)^{1/v} \right] = -\log \pi_{ijm}^g\end{aligned}$$

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 \bar{W}^g , and the final uses the choice

obtain $\Pi \equiv \sum_{i,j} \Pi_{ij}$. Finally, substituting each element into (12), using the mode-specific definitions of commute utility U_{ijm}^g , and rearranging, I obtain the expression in (C.1). \square

C.2 Decentralization of Social Planner Optimum

In this section, I derive the per-minibus subsidies to associations z_{ij} as well as per-minibus-commuter subsidies t_{ij} that induce the socially-optimal bus supply b_{ij}^* and commute choices π_{ijm}^{g*} under full association bargaining power, $\beta = 1$, and, as always, Assumption 1. First, consider bus supply. Under the latter condition, associations freely choose fares τ_{ij} and b_{ij} on their route to maximize profits, such that fares are determined by the first-order condition

$$\frac{\partial \Pi_{ij}}{\partial \tau_{ij}} = \left[\tau_{ij} - \frac{\chi_{ij}}{\bar{\eta}} - \frac{\bar{\omega}}{\bar{\eta}} (L_{ij} + T_{ij}) \right] \sum_g N^g \frac{\partial \pi_{ijM}^g}{\partial \tau_{ij}} + \frac{\lambda_{ij}}{\bar{\eta}} \left[\bar{\eta} - \bar{\omega} \frac{\partial L_{ij}}{\partial \lambda_{ij}} \sum_g N^g \frac{\partial \pi_{ijM}^g}{\partial \tau_{ij}} \right] = 0 \quad (\text{C.3})$$

and thus satisfy

$$\tau_{ij} = \frac{\chi_{ij}}{\bar{\eta}} + \frac{\bar{\omega}}{\bar{\eta}} (L_{ij} + T_{ij}) - \frac{\lambda_{ij}}{\sum_g N^g \frac{\partial \pi_{ijM}^g}{\partial \tau_{ij}}} + \frac{\lambda_{ij} \bar{\omega}}{\bar{\eta}} \frac{\partial L_{ij}}{\partial \lambda_{ij}} \quad (\text{C.4})$$

With a per-bus flow subsidy of z_{ij} on route ij , associations' first-order condition (A.7) from Appendix A.3, after substituting in optimally-chosen fares (C.4), becomes

$$\begin{aligned} & \left\{ -\frac{\lambda_{ij}}{\sum_g N^g \frac{\partial \pi_{ijM}^g}{\partial \tau_{ij}}} + \frac{\lambda_{ij} \bar{\omega}}{\bar{\eta}} \frac{\partial L_{ij}}{\partial \lambda_{ij}} \right\} \sum_g N^g \frac{\partial \pi_{ijM}^g}{\partial b_{ij}} \\ & \quad - \frac{\lambda_{ij}}{\bar{\eta}} \bar{\omega} \left(\frac{\partial L_{ij}}{\partial \lambda_{ij}} \sum_g N^g \frac{\partial \pi_{ijM}^g}{\partial b_{ij}} + \frac{\partial L_{ij}}{\partial b_{ij}} \right) - \bar{\omega} + z_{ij} = 0, \end{aligned} \quad (\text{C.5})$$

which coincides with the optimal entry b_{ij}^* pinned down by the planner's bus entry condition (A.5) from Appendix A.3 if and only if the per-bus-supplied subsidy satisfies

$$z_{ij} = - \sum_g N^g \pi_{ijM}^{g*} r \omega_j^g \left[\frac{\partial Q_{ij}}{\partial b_{ij}} + \frac{\partial L_{ij}}{\partial b_{ij}} \right] + \frac{\lambda_{ij}^*}{\sum_g N^g \frac{\partial \pi_{ijM}^g}{\partial \tau_{ij}}} \sum_g N^g \frac{\partial \pi_{ijM}^g}{\partial b_{ij}}. \quad (\text{C.6})$$

where λ_{ij} indicates the passenger arrival rate under optimal commuter choices.

Second, I derive commuter subsidies t_{ij} to minibus commuters differentiated by home and work which induce the optimal choices π_{ijm}^{g*} . Subsidies t_{ij} are paid along with the fare upon choice of a commute. By comparing the social planner-optimal choice probabilities (A.6) from Appendix A.3

probability equation (8).

with the equilibrium (relative) choice probabilities implied by (8) and substituting in minibus fares (C.4), we can see that the optimal choice probabilities can be implemented by subsidies according to

$$t_{ij} = -\frac{\lambda_{ij}^*}{\sum_g N^g \frac{\partial \pi_{ijM}^g}{\partial \tau_{ij}}} - \sum_g N^g \pi_{ijM}^{g*} \left[r \omega_j^g \left(\frac{\partial Q_{ij}}{\partial \lambda_{ij}} + \frac{\partial L_{ij}}{\partial \lambda_{ij}} \right) \right]. \quad (\text{C.7})$$

In welfare calculations, I assume that commuters of a given skill pay equal lump-sum taxes to cover the costs of (i) the commuter subsidies paid out to *their own skill group* and (ii) a share of bus supply subsidies equal to the skill group's aggregate population share.

C.3 Micro-Foundations of Commute Utility

I now lay out a micro-founded model of commuting which underlies the linear approximations to commute utility U_{ijm}^g used in the main text. Suppose that commuters, upon birth, immediately enjoy their home-location-specific amenity θ_i^g , pay a one-time mode-specific utility cost κ_M^g and then wait for their chosen mode m , so their total (deterministic) utility satisfies $\bar{U}_{ijm}^g = \theta_i^g - \kappa_M^g + E[u_{ijm}^g]$ 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_{ijm}^g , satisfies

$$u_{ijm}^g = \exp(-rw_{ijm}) (V_{ijm}^g - \tau_{ijm}), \quad (\text{C.8})$$

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

$$V_{ijm}^g = \exp(-rt_{ijm}) \omega_j^g, \quad (\text{C.9})$$

Substituting (C.9) into (C.8) and taking a first-order approximation, around $r = 0$, to the value of waiting yields

$$u_{ijm}^g \approx \omega_j^g - \tau_{ijm} - r \omega_j^g (w_{ijm} + t_{ijm}) + r \tau_{ijm} w_{ijm} \quad (\text{C.10})$$

Finally, taking expectations and substituting into total utility deterministic utility \bar{U}_{ijm}^g yields

$$\bar{U}_{ijm}^g \approx \underbrace{\theta_i^g - \kappa_m^g - r \omega_j^g [E(w_{ijm}) + t_{ijm}] - \tau_{ijm}}_{\equiv U_{ijm}^g} + \omega_j^g. \quad (\text{C.11})$$

where I have suppressed the term $r \tau_{ijm} w_{ijm}$, which will be close to zero due to the multiplication of

a small time preference rate and a small fare. I then define commute utility U_{ijm}^g as the component corresponding to the costs of the actual commute. For the minibus mode, $m = M$, I set $E(w_{ijM}) = Q_{ij} + L_{ij}$, $t_{ijM} = T_{ij}$, and $\tau_{ijM} = \tau_{ij}$ to obtain Equation (7) in the main text. For formal transit, $m = F$, $E(w_{ijF}) = H_{ij}$ and $t_{ijF} = T_{ijF}$, while, for the car mode, $m = A$, $E(w_{ijA}) = 0$, $t_{ijA} = T_{ij}$, and $\tau_{ijm} = \tau_A$.

D. ESTIMATION

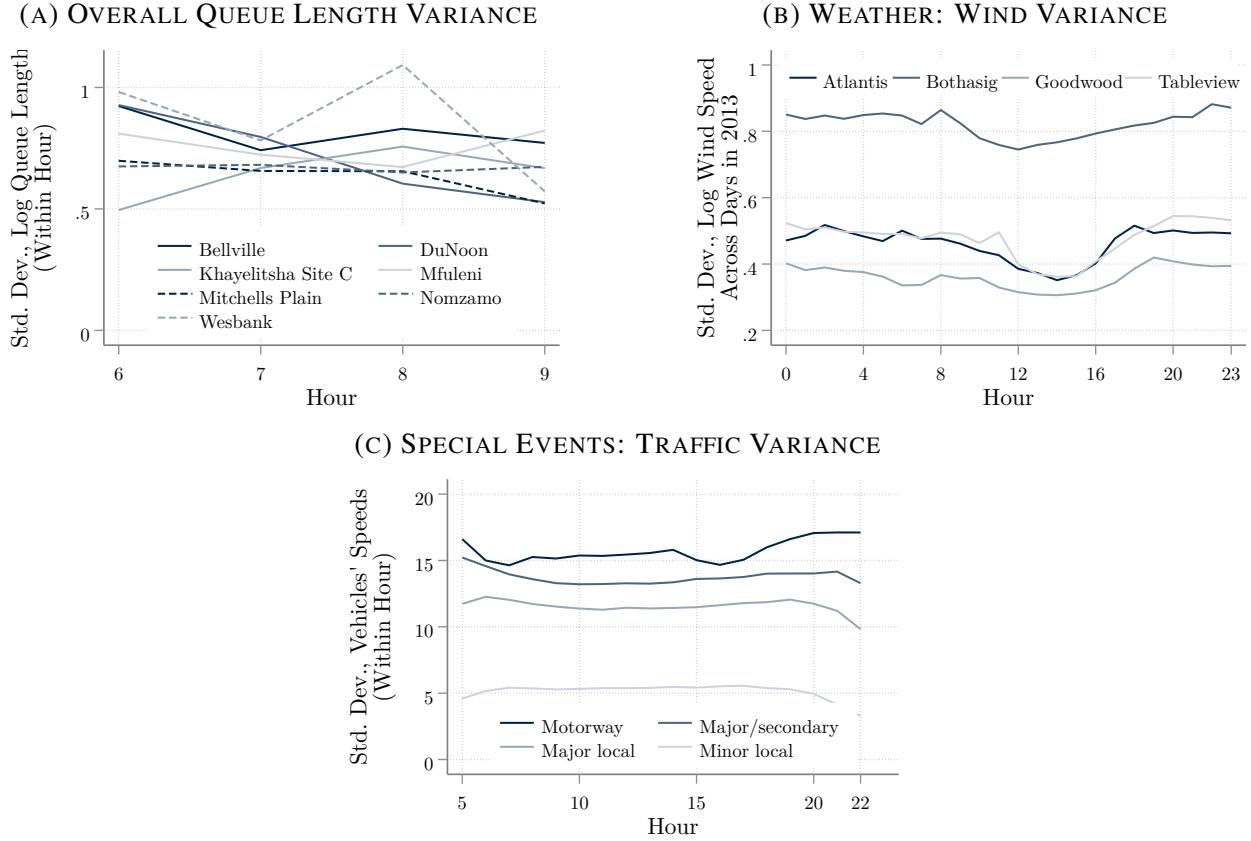
D.1 Robustness

D.1.1 Queueing Efficiency

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 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 over the course of the 6-10am morning commute. In sum, available evidence suggests that key sources of minibus loading delays vary little

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



Notes: Panel (A) displays the standard deviation of log queue lengths across 5-min.-time 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.

over the course of the morning rush hour, so any correlation with my work start times instrument must be mild.

D.1.2 Stated Preference

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 two ways. First, I re-estimate the model using the city-conducted survey plus only those respondents in my survey interviewed at the Middestad Mall/Bellville intermodal interchange, a recruitment location less prone to oversampling of minibus riders. Column 2 of Online Appendix Table D.1 shows that the estimated parameters, though somewhat noisier, mirror my full-sample estimates quite closely, in particular the rate of time

TABLE D.1. STATED PREFERENCE: ROBUSTNESS TO SAMPLE

Parameter	Skill	(1)	(2)	(3)
		Baseline	Intermodal Sample Only	Commute Mode-Weighted
r		0.001 (0.0004)	0.0014 (0.0007)	0.0011 (.0005)
v		4.76 (1.26)	6.83 (2.73)	5.84 (1.99)
κ_M	<i>Low</i>	7.68 (1.56)	10.61 (3.54)	9.25 (2.55)
	<i>High</i>	15.03 (3.55)	21.16 (7.82)	18.3 (5.67)
ξ_{security}	<i>Low</i>	-1.09 (0.39)	-2.13 (1.06)	-1.55 (0.69)
	<i>High</i>	-2.75 (0.84)	-4.91 (2.29)	-5.1 (1.86)
$\xi_{\text{no overloading}}$	<i>Low</i>	-1.38 (0.437)	-2.02 (1.01)	-1.26 (0.596)
	<i>High</i>	-1.39 (0.543)	-1.25 (1.28)	-1.43 (0.83)
$\xi_{\text{no speeding}}$	<i>Low</i>	-1.36 (0.44)	-3.03 (1.38)	-2.12 (0.85)
	<i>High</i>	-0.825 (0.465)	-1.86 (1.39)	-0.582 (0.73)
κ_F	<i>Low</i>	3.63 (0.51)	4.53 (1.08)	4.14 (0.80)
	<i>High</i>	9.17 (1.89)	12.5 (4.20)	10.96 (3.05)
<i>N</i> Respondents		820	546	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 (16). 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 the Middestad Mall/Bellville transport interchange (i.e. excluding those sampled at minibus stations); Column 3 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.

preference r , Gumbel scale v , and the high value the high-skill place on minibus station security. Even this “intermodal sample,” however, still oversamples minibus commuters. Thus, in Column 3, I weight my own sample, which, critically, does not contribute to identification of the 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 takeways remain.

D.1.3 Minibus Arrivals

As discussed in Section V, I estimate a model-implied equation for the (average) gap, $\overline{\text{bus gap}}_{ijt}$, in minutes between the departure of one bus from the origin station and the arrival of the next,

$$\log(\overline{\text{bus gap}}_{ijt}) = -\log \alpha_0 - \alpha_1 \log b_{ijt} + \vartheta_{ijt}, \quad (\text{D.1})$$

across routes ij and hours t , as a function of the number of buses operating on a route, b_{ijt} , in my station count data. The unobserved arrivals efficiency, $\vartheta_{ijt} \equiv -\log \varsigma_{ijt}$, 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-hour fixed effects. To address any differential impacts of, say, foot traffic on different routes operating out of the same origin station, I instrument for bus supply on a route with the distance from that route's origin to destination distance. 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.2, I find a bus arrivals elasticity of $\hat{\alpha}_1 = 0.38$. When I add origin-hour fixed effects in Column (2), the magnitude of the elasticity rises, but the Column (1) estimate remains within the corresponding 95% confidence interval. Finally, in Column (3), I find a qualitatively similar elasticity when I employ the distance instrument. However, the instrument proves relatively weak. Thus, to obtain a transparent estimate of the intercept α_0 , I employ the Column (1) estimates in the model quantification and counterfactuals in Sections VI-VII.

D.2 Externally Calibrated Parameters

D.2.1 Geography

The model geography consists of the $I = 18$ transport analysis zones (TAZ) in Cape Town. I exclude from commuters' choice sets one home location that is, in reality, an essentially non-residential TAZ with less than 5,000 residents employed within Cape Town. 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.

TABLE D.2. MINIBUS ARRIVALS FUNCTION ESTIMATES

	OLS		IV
	(1) log bus gap	(2) log bus gap	(3) log bus gap
α_0	0.24		
<i>Arrivals Intercept</i>	(0.10)		
α_1	0.38	0.59	0.73
<i>Arrivals Elasticity</i>	(0.16)	(0.17)	(0.30)
95% CI for α_1 :		[0.19,1.00]	[0.01,1.44]
Weak IV-Robust P-Value for $\alpha_1 = 0$:			0.27
Origin-by-hour FE		✓	✓
Observations	161	159	157
Adjusted R^2	0.09	0.32	
First-Stage F Statistic			2.42

Notes: Robust standard errors in parentheses, clustered at the origin level. Table presents estimates of (D.1) over hours and 44 routes in my station count data, with origin by hour fixed effects included, as noted. In Column 3, I instrument for bus supply using the log straight-line distance from a route's origin to destination. For the IV specification, I additionally report the effective first-stage F statistic following Olea and Pflueger (2013) as well as the p-value associated with the Anderson-Rubin chi-squared test robust to weak instruments (Andrews et al. (2019)).

D.2.2 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 income of workers in a skill group employed outside the home.³⁹

³⁹To 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 also fares, to USD for scaling purposes.

D.2.3 On-Road Travel Time T_{ij} and Formal Transit Wait and Travel Times H_{ij}, T_{ijF}

I then calibrate the on-road travel time T_{ij} as well as the formal transit wait, H_{ij} and travel times, T_{ijF} using stylized transport networks specific to each mode. The nodes of the road network consist of the 18 TAZ with links connecting each pair of TAZ which share a border, while the formal transit network consists of (potential) connections between every pair of locations. The Microsoft Azure API provides on-road driving times as well as average formal transit wait and travel times along each link of the respective networks.⁴⁰ For either the road or formal transit network and a given origin-destination pair ij , I calculate the total wait time, if applicable, to obtain H_{ij} , and the total travel time, to obtain T_{ij} and T_{ijF} , along the shortest path through the network.

D.2.4 Formal Transit Fares τ_{ijF} and Car Commute Cost τ_A

Formal transit 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.⁴¹

D.2.5 Minibus Operations Parameters $\bar{\eta}, \chi_{ij}$

I set minibus capacity to the $\bar{\eta} = 15$ 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 proportionate to driving distance and then use fuel efficiency figures provided by the firm GoMetro to calibrate the per-kilometer cost.⁴²

D.2.6 Emissions Parameters χ_m^e and ς

I obtain mode-specific carbon-equivalent emissions from calculations in Borck (2019) combined with U.S. Department of Energy estimates. Specifically, I take Borck (2019)’s estimate of $\chi_A^e = 0.554$ kg CO₂-equivalent emissions per kilometer from driving, a figure which includes actual

⁴⁰I calibrate the driving time along each link by querying the Microsoft Azure API (similar to Google Maps) for a trip between the centers of employed population of the two TAZ on a Wednesday at 7am. From the Azure API, I also obtain formal transit wait and travel time for each link in the transit network from averages over 6 evenly-spaced trips on a Wednesday between 7:00 and 8:00am via formal transit modes: Metrorail commuter trains, Golden Arrow private scheduled buses, and MyCiti bus rapid transit. The formal transit wait time equals the time between the queried departure time and the actual departure time of the suggested itinerary.

⁴¹The 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.

⁴²I use the driving distance under free-flow (Sunday, 11pm) conditions between 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.

CO₂ emissions as well as “local pollutant” emissions. To obtain emissions from other modes, I use relative passenger-mile-per-gallon (pmpg) estimates from my partner firm GoMetro and the Alternative Fuels Data Center (U.S. Department of Energy (2023)).⁴³ For the social cost of carbon, I follow Borck (2019)’s benchmark of $\zeta = \$0.0485$ per kilogram CO₂. The per-commuter emissions cost on mode m from home i to work j then equals $E_{ijm} \equiv \zeta \chi_m^e \times \text{driving distance}_{ij}$, where I calculate driving distance as in Online Appendix D.2.5.

E. VALIDATION AND RESULTS

E.1 Mode Shares

In this section, I compare actual commute mode shares to those predicted by my model. In particular, 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.

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 (16) and interact a dummy variable for the binary demographic characteristics listed in Column 1 of Online Appendix Table E.2 with the terms in the multinomial logit model that identify utility costs, their dependence on policy, and the value of time. Women and college workers have a higher value of time saved, even conditional on income; the 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.

⁴³Fuel efficiency estimates provided by GoMetro suggest that the most common minibus vehicle, the gasoline-powered Toyota Quantum, requires 0.143 liters/km, equivalent to 248.05 passenger miles per gallon with a full load of 15 passengers. The AFDC estimates an average of 27.5pmpg for single-occupancy cars, so minibuses have 0.11 the energy use of cars per passenger-distance. Applying this factor to χ_A^e yields $\chi_M^e = 0.0615$. The AFDC estimates 137.2pmpg for high-ridership bus and 600pmpg for high-ridership train. Taking a weighted average of these using the 53% share of formal transit commuters in the 2013 Cape Town Household Travel Survey who use trains at some point during their commutes, I obtain a formal transit average of 382.48pmpg. Thus, formal transit has 0.07 the fuel use of cars, yielding $\chi_F^e = 0.0388$.

TABLE E.1. MODE CHOICE PROBABILITIES, DATA VS. MODEL

Variables	Minibus	Car
	Mode Share, Data	Mode Share, Data
Mode Share, Model	0.980 (0.129)	1.157 (0.0947)
Constant	-0.0744 (0.0300)	0.281 (0.0301)
Observations	497	497
R-squared	0.082	0.232

Robust standard errors in parentheses

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. STATED PREFERENCE HETEROGENEITY: DIFFERENCE IN PARAMETER ESTIMATE, VERSUS BASE CATEGORY

Dimension	r	Mode Utility Cost		Effects on Minibus Utility Cost		
		κ_M	κ_F	$ \xi_{\text{overload}} $	$ \xi_{\text{security}} $	$ \xi_{\text{speed}} $
Female	0.0013 (0.0006)	-3.61 (1.06)	-3.27 (0.924)	-0.222 (0.419)	-1.33 (0.535)	-0.49 (0.436)
College	0.0019 (0.0007)	6.66 (1.94)	4.62 (1.28)	0.052 (0.481)	1.71 (0.659)	-0.458 (0.499)
Age>45	0.0027 (0.001)	-1.03 (0.709)	-1.80 (0.671)	0.494 (0.640)	1.72 (0.770)	2.50 (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 (16). 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 H.H. Travel Survey stated pref. modules.

E.3 Counterfactual Robustness

In this section, I show how alternative modeling assumptions alter the welfare gains from the same optimal minibus and commuter subsidies simulated in Section VII. I re-solve for the baseline equilibrium under each set of alternative assumptions. For the sake of comparison, I do not

recalculate the social optimum but rather simulate exactly identical subsidy levels, again under the new set of modeling assumptions.

Road Congestion

First, I consider how road congestion might alter the welfare gains from optimal minibus and commuter subsidies. I employ the stylized road network described in Online Appendix D.2.3. The travel time over an individual link between two adjacent transport analysis zones, $t_{ik} \equiv \bar{t}_{ik} v_{ik}^\gamma$, now depends on a calibrated intercept \bar{t}_{ik} and increases with the vehicle inflow v_{ik} , raised to a road congestion elasticity γ . For tractability, I require that minibuses and cars follow the sequence of links that minimizes free-flow travel time and calculate T_{ij} as the sum of congestion-affected travel times t_{ik} over links traversed.

In separate results available upon request, I employ road-segment-level data for Cape Town from TomTom's Traffic Stats API to estimate $\hat{\gamma} = 0.0917$, closely tracking the baseline instrumental variables estimate of the equivalent log-linear effect of traffic volume on urban-segment travel time in Allen and Arkolakis (2022). I then calibrate \bar{t}_{ij} to match observed driving times.

Minibuses make up only a small share of overall traffic, so the gains from the subsidies remain almost unchanged, as Online Appendix Table E.3 enumerates. Nevertheless, both skill groups benefit slightly more from the social optimum due to associated reductions in (car) traffic.

Nested Logit

Second, 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

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

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

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

The home-work-mode location choice probabilities then satisfy

$$\pi_{ijm}^g \equiv \pi_{ij}^g \pi_{m|ij}^g. \quad (\text{E.3})$$

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

I set $v = 4.76$ 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 $\zeta = 5.97$.⁴⁴

This nested logit demand system, as calibrated, now means commuters less willingly change home or work locations than they do modes. Thus, the decrease in earned wages in Online Appendix Table E.3 is somewhat milder than in the main text, but commuters also shift less strongly towards the routes with the largest wait time decreases, so overall welfare gains from the optimal subsidies are only 60-70% as high as in the baseline, non-nested case but nonetheless remain comparable to the gains estimated for other transport investments in the literature.

⁴⁴Specifically, let $\tilde{\zeta}$ given Frechet elasticity. Denote the expected utility of home-work pair ij by $\bar{\Omega}_{ij} \equiv \theta_i^g + \bar{W}_{ij}^g + \omega_j^g$; then, the Frechet elasticity satisfies $\tilde{\zeta} = \frac{\partial \log(\pi_{ij}/\pi_{kl})}{\partial \log \bar{\Omega}_{ij}}$. In my logit demand system, instead,

$$\frac{1}{\zeta} = \frac{\partial \log(\pi_{ij}/\pi_{kl})}{\partial \bar{\Omega}_{ij}} = \frac{\partial \log(\pi_{ij}/\pi_{kl})}{\partial \log \bar{\Omega}_{ij}} \frac{1}{\bar{\Omega}_{ij}}$$

where the final equality uses the chain rule. Thus, a parameter ζ in my model equivalent to the Frechet parameter $\tilde{\zeta}$ must satisfy $\zeta = \frac{\bar{\Omega}_{ij}}{\tilde{\zeta}}$. I calculate the mean realized value of $\bar{\Omega}_{ij}$, weighted by choice probabilities, in my baseline model, and divide by $\tilde{\zeta} = \eta_l = 2.07$ from Tsivanidis (2023) to obtain $\zeta = 5.97$.

Endogenous Bus Departures

Third, I model and quantify the implications of associations' choice of the number of passengers η_{ij} each bus loads, on average, before departure, up to bus capacity $\bar{\eta}$. As in the main text, I employ a probabilistic structure whereby, given the *departure threshold* η_{ij} , a bus departs at rate $\tilde{\lambda}_{ij}/\eta_{ij}$, where $\tilde{\lambda}_{ij}$ denotes the expected outflow of passengers from the queue onto the bus.

Conditional on bargained fares, the association on each route then solves

$$\max_{b_{ij}, \eta_{ij} \leq \bar{\eta}} \left\{ \frac{\lambda_{ij}}{\eta_{ij}} [\eta_{ij} \tau_{ij} - \chi_{ij} - \bar{\omega} (L_{ij} + T_{ij})] - \bar{\omega} b_{ij} \right\}.$$

Associations' first-order condition for the departure threshold,

$$0 = \frac{\partial \Pi_{ij}}{\partial \eta_{ij}} = \frac{\Pi_{ij} + \bar{\omega} b_{ij}}{\lambda_{ij}} \frac{\partial \lambda_{ij}}{\partial \eta_{ij}} + \lambda_{ij} \left[\frac{\chi_{ij}}{\eta_{ij}^2} + \frac{\bar{\omega} (L_{ij} + T_{ij})}{\eta_{ij}^2} \right] - \frac{\lambda_{ij}}{\bar{\eta}} \bar{\omega} \left(\frac{\partial L_{ij}}{\partial \lambda_{ij}} \frac{\partial \lambda_{ij}}{\partial \eta_{ij}} + \frac{\partial L_{ij}}{\partial \eta_{ij}} \right), \quad (\text{E.4})$$

highlights the tradeoffs at work in associations' decision of whether to allow buses to depart with a larger number of passengers. In particular, the first term represents profits lost due to the loss in demand from longer loading times, and the second reflects the fact that taking on more passengers allows the fixed operations and labor costs of a trip to be "spread" across more fare-paying commuters. The final term captures the additional labor costs of longer loading times, adjusted for changes in demand, which themselves affect loading times. The driver wage $\bar{\omega}$ proves a key quantity which determines the magnitude of the "fixed" labor costs of each trip and thus increases associations' incentives to load more passengers per bus.

For the sake of brevity, I avoid restating the equilibrium conditions and formal equilibrium definition. Instead, equilibrium now consists of a vector $\{b, \tau, \pi, \lambda, P^L, Q, L, \eta\}$ that satisfies (i) (1)-(3), (5), (6), (8), and (9) with η_{ij} always in place of $\bar{\eta}$; (ii) the departure threshold first-order conditions (E.4); and (iii) the bus capacity constraints $\eta_{ij} \leq \bar{\eta}$.

I recalculate the effects of the same optimal minibus and commuter subsidies from the main text in this new model. In the baseline equilibrium with endogenous departures, 99.3% of routes depart full, i.e. with $\omega_{ij} = \bar{\eta}$; the planner-induced bus entry slows loading on the lowest-demand routes, so the share departing full falls to 81.3%. The somewhat shorter loading times on these mostly low-wage, low-demand routes contribute to modestly higher welfare gains in the third line of Online Appendix Table E.3. However, the effects on mode shares, wages, emissions, and welfare closely track those with all buses departing "full."

TABLE E.3. ROBUSTNESS: SOCIAL PLANNER COUNTERFACTUAL

Policy	Skill:	Change in Mode Share				% Change in...					
		Minibus		Car		Earned Wage	Emissions	Welfare		Welfare, Net of Emissions	
		Low	High	Low	High			Low	High	Low	High
Road Congestion		0.03	0.03	-0.01	-0.02	-0.32	-2.12	1.36	0.41	1.39	0.42
Nested Logit		0.03	0.03	-0.01	-0.02	-0.23	-2.33	0.80	0.27	0.79	0.26
Endogenous Bus Departures		0.03	0.03	-0.01	-0.02	-0.32	-2.14	1.35	0.41	1.38	0.42

Notes: This table summarizes the effects of the implementation of the social optimum via optimal minibus and commuter subsidies, under three variations on the model in Section IV in the main text: (i) road congestion effects of minibus and car traffic on the travel times of minibuses and cars; (ii) nested logit commuter demand, with a nest for each home and work location pair comprising the three modes; (iii) associations choose with how many passengers, on average, to depart. The first four columns show the changes in the minibus and car mode shares by skill group. The fifth and sixth show the percent changes in the average wage *earned* by commuters, gross of commute costs, and total emissions, which I calculate as described in Online Appendix D.2.6. The final four columns show the percent change in group-level welfare, measured as equivalent variation and, in the last two columns, net of external emissions costs.

E.4 Additional Policy Counterfactuals

I now explore three additional minibus-targeted policy counterfactuals, detailed in Online Appendix Table E.4.

Two Queues Only on Top 10% of Routes by Demand

First, since all routes except those with the highest demand lost from the across-the-board second-queue counterfactual in the main text, I implement a more targeted policy: only the routes with baseline equilibrium passenger inflows, λ_{ij} , above the 90th percentile across routes receive a second queue. These busiest routes benefit from increasingly large decreases in total queueing plus loading time. Other routes' total waits remain essentially unchanged, versus the mild increases these routes experienced upon receipt of a second queue in the main text. As a result, each skill group, in Online Appendix Table E.4, gains a qualitatively similar but marginally greater amount in equivalent variation terms, versus the blanket second queue case in Table 5.

Smaller Minibuses Only on Bottom 20% of Routes by Demand

Second, in a similar vein, I implement the smaller 12-person buses only on the 20% of routes with lowest baseline demand, λ_{ij} . These routes, given their long loading times, stand to benefit most; limiting the policy to the lowest-demand routes, in the second line of Online Appendix Table E.4, does eliminate the queueing time increases experienced by the busiest routes and thus the slight losses from the across-the-board smaller bus size. Nonetheless, the routes which receive

TABLE E.4. ADDITIONAL COUNTERFACTUAL URBAN TRANSPORTATION POLICIES

Policy	Skill:	Change in Mode Share				% Change in...					
		Minibus		Car		Earned Wage	Emissions	Welfare		Welfare, Net of Emissions	
		Low	High	Low	High			Low	High	Low	High
Two Queues [Top 10% by Demand]		0.03	0.01	-0.01	-0.01	0.83	-1.81	1.70	0.37	1.73	0.38
Smaller Minibuses [Bottom 20% by Demand]		0	0	0	0	0	0	0	0	0	0
Lower Driver Wage		0	0	0	0	0.06	-0.16	0.15	0.04	0.15	0.04

Notes: Table summarizes the effects of additional counterfactuals. The first two, adding a second queue to routes above the 90th percentile of baseline demand λ_{ij} and reducing minibus capacity $\bar{\eta}$ to 12 on routes below the 20th percentile of baseline demand, use the baseline model in Section IV in the main text. The third, decreasing the driver wage $\bar{\omega}$ by 50%, uses the model with endogenous bus departures outlined in Online Appendix E.3. The first four columns show the changes in the minibus and car mode shares by skill group. The fifth and sixth show the percent changes in the average wage *earned* by commuters, gross of commute costs, and total emissions, which I calculate as described in Online Appendix D.2.6. The final four columns show the percent change in group-level welfare, measured as equivalent variation and, in the last two columns, net of external emissions costs.

smaller buses and thereby benefit from lower queueing and loading times serve, by construction, very few passengers. Higher fares prevent additional commuters from switching to these routes, so the targeted smaller buses have only minuscule welfare effects. Recall that, under the admittedly strong but not highly consequential assumption that operations costs χ_{ij} do not depend on bus size, this counterfactual is again isomorphic to requiring the original 15-passenger buses to depart, on average, once they reach 12 passengers.

Lower Driver Wage with Endogenous Bus Departures

Third, I consider the effect of increasing minibus driver wages $\bar{\omega}$ in the model from Online Appendix E.3 where associations choose with how many passengers, η_{ij} , to depart. Associations, faced with a smaller fixed labor cost per bus trip, might theoretically choose a lower η_{ij} to decrease loading times and attract additional demand. However, even halving the driver wage is not sufficient to cause early departures: as in the baseline, buses on 99.3% of routes depart full. The lower driver wage does provoke associations on the busiest, highest- λ_{ij} routes to increase bus supply, so the lower queueing times which ensue contribute to modest welfare gains.