

# Privatized Provision of Public Transit\*

Lucas Conwell <sup>†</sup>

First Version: October 2022

This Version: April 2025

## Abstract

Privately-operated minibuses provide 50–100% of urban transit in often cash-strapped developing-country cities, at the cost of long wait times and poor personal safety. To quantify the gains from low-cost reorganization of this privatized provision of public transit, I build the first model of privatized shared transit, which features increasing returns in the form of shorter waits on busier routes and a key role for surplus bus supply as “insurance” against demand spikes. Market power of local minibus associations inhibits realization of the former but facilitates internalization of the latter. I then estimate the model with newly-collected data on minibus arrivals and passenger queues in Cape Town as well as stated preferences for exogenously varied commute attributes. A formalization program of government-set fares and subsidies leverages increasing returns to shrink wait times and queues, while government actions to enforce speed limits 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 Gharad Bryan, Nina Harari, Allan Hsiao, Gabriel Kreindler, Oluchi Mbonu, Melanie Morten, Nick Tsivanidis, and Román David Zárate for their insightful comments and suggestions. I am also grateful to Treb Allen, Richard Blundell, Richard Carson, Alvaro Cox, Aureo de Paula, Mateus Dias, Fabian Eckert, Simon Fuchs, Jonas Hjort, Imran Rasul, Edouard Schaal, and Gabriel Ulyssea for their feedback which materially improved the paper. I am also grateful to the anonymous referees who provided useful and detailed comments on an earlier version of the manuscript. I thank Philip Krause, Aslove Mateyisi, and Mokgadi Mehlape from GoAscendal/GoMetro for their herculean efforts to obtain permission for and organize the data collection. Finally, I acknowledge generous financial support from the Stone Centre at UCL, the Yale Economic Growth Center, and the Ryoichi Sasakawa Young Leaders Fellowship Fund that made this project possible. IRB Exemption Determination obtained through Yale University, ID 2000032302.

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

# I. INTRODUCTION

In fiscally-constrained developing-country cities, the private sector provides many of the urban public goods which remain the sole preserve of the state in the developed world (Bryan et al. (2020)). In the quintessential case of transportation, for example, limited resources and state capacity frequently preclude investment in costly formal public transit infrastructure (Balboni et al. 2020). As a result, local workers waste significant time commuting (Kreindler (2022)), and these often-unsafe journeys constrain their access to productive jobs (Tsivanidis (2023) and Zarate (2024)), which might, in turn, bolster urban incomes. 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)).

The privatized provision of transit, however, comes with stark tradeoffs. On the one hand, minibuses, with only minimal public investment, knit together even the farthest-flung, lowest-income suburbs and profitably operate scores of low-demand routes. On the other hand, even to the most casual observer, the downsides of these privatized networks loom large: unsafe driving, congested roads, and, chief among them, long wait times for passengers. Wait time alone accounts for a full one-third of the average public transit commute time worldwide (Nikolaidou et al. (2023)).<sup>1</sup> Indeed, in Cape Town, a quintessentially sprawling, privatized-transit dominated sub-Saharan African city, wait time consumes 15 minutes, or again over a third, of the average minibus trip. Perhaps unsurprisingly, then, local policymakers typically view minibuses as a hindrance to development – a view which economists, in their focus on formal, government-provided alternatives, have rarely challenged. Nonetheless, these networks’ very existence raises a tempting prospect for fiscally-constrained governments across the developing world. Could cities, via low-cost interventions, improve upon these existing *privatized shared transit* networks to boost urban connectivity and ultimately lift more residents out of poverty?

In this paper, I answer this question by introducing what is, to my knowledge, the first model of privatized shared transit. Existing urban models typically feature exogenous commute costs and thus cannot characterize how, for example, wait times or fares respond to policy. My model, in contrast, endogenizes both the time and monetary costs of minibus commutes. Higher commuter demand and bus supply decrease passenger waits, which generates a form of increasing returns, and the equilibrium responses of wait times, road congestion, and fares interact with the spatial sorting of commuters. Virtually no data exists on these wait times or on bus supply, so I collected new data on passenger queues and bus arrivals in Cape Town. To quantify commuters’ preferences for difficult-to-measure minibus attributes, e.g. security guards, I conducted discrete choice experiments which circumvent both incentive problems and the identification issues inherent in the usual observational data. With the estimated model, I evaluate prototypical policies which require minimal upfront investment yet substantially raise welfare simply by alleviating minibus-sector frictions. An optimal minibus formalization program would lower fares to fill buses faster and thus decrease wait times, while simultaneously subsidizing surplus bus supply to prevent the build-up of queues. Speed limit enforcement or government-provided station security guards, in turn, generate even larger welfare and emissions gains.

---

<sup>1</sup>A long line of work in urban planning discusses the costs of wait time for public (Abenoza et al. (2019), Ansari Esfeh et al. (2021), and De Vos et al. (2023)) as well as privatized transit (Cervero and Golub (2007) and Kerzhner (2022)).

I begin with four facts about the minibus market in Cape Town. First, wait time accounts for 36% of total commute time on the average minibus route – and particularly plagues suburban routes. Second, passengers wait less on busier minibus routes, both in queues until a bus arrives and on board the bus, the latter because buses emulate the *fill-and-go* practice virtually universal among privatized shared transit providers worldwide. Third, passenger arrival shocks lead to the build-up of queues, but additional bus supply acts as “insurance” and dampens the growth of queues. Fourth, among all attributes of minibuses, commuters express the highest levels of dissatisfaction with road safety and security from crime.

I then start from first principles to build a new model of the privatized shared transit sector, which, unlike the fixed “iceberg” commute costs typical in the literature, captures crucial equilibrium responses of wait times and fares. In my model, a single minibus association on each distinct origin-destination route chooses the supply of buses and Nash-bargains fares with the city government, a setup which grants these associations a form of limited market power. A queueing model, tailored to the minibus loading process and solved out of steady-state, determines passengers’ wait times. Passengers first queue until a minibus arrives and subsequently wait on these buses, which depart only when full. Travel times, in turn, increase with road traffic.<sup>2</sup> 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.

My model highlights three potential sources of inefficiency. First, minibuses’ strict capacity constraint bakes in a negative relationship between demand and passenger wait times: twice as many passengers require twice-as-frequent departures. I term this source of increasing returns the *Market Mohring Effect*, in reference to the related economies of scale from which all forms of shared transit benefit (Mohring (1972)). Second, the constrained size of minibuses also introduces the potential for the build-up of passenger queues after sudden spikes in demand. Surplus bus supply helps dampen queues, in a novel *Bus Supply Insurance Effect* unique to low-capacity minibuses. Third, minibuses and car commuters impose the standard road congestion externalities on fellow road users. When it comes to efficiency, market power in privatized shared transit is a double-edged sword. On the one hand, greater bargaining power allows associations to charge higher fares, which suppresses demand and thereby inhibits the realization of the Market-Mohring increasing returns. On the other hand, higher fares induce associations to internalize more of the gains from additional bus supply “insurance.”

I collected two forms of primary data in Cape Town to directly measure wait times and quantify consumer preferences for typically-unobserved commute attributes relevant to the aforementioned minibus reforms. First, enumerators observed passenger queues and bus arrivals on a random sample of 44 minibus routes, from which I calculate wait times. Second, I conducted commute mode discrete choice experiments, which involve a tangible, familiar decision particularly suited to such stated preference methods. In my survey, 526 respondents chose among hypothetical minibus commute options with exogenously varied travel times, costs, and quality improvements, e.g. security guards. Preferences across modes, not only minibuses but also car and “formal” public transit, come from a separate city-run stated preference survey.

---

<sup>2</sup>I abstract away from stops en route, which Cape Town’s highway-reliant road network largely prevents.

I then estimate the model’s structural parameters in four main steps. First, I directly estimate passenger queueing efficiency from the relationship between the prevalence of queues and the time buses take to fill up. Unobserved interruptions to this loading process, e.g. due to weather or special events, could lengthen loading times, so I instrument for year-2022 queues with the times commuters report leaving their homes in a 2013 survey, likely uncorrelated with interruptions nine years later. I also estimate a bus arrival efficiency parameter from the effect of bus supply on the gaps in time between successive buses, where I instrument for bus supply with route distance. Second, the discrete choice experiments provide exogenous variation that identifies commuters’ mode-specific utility costs as well as values of time, driving safety, and security in dollar equivalents. Third, I estimate the road congestion elasticity from TomTom traffic data and externally calibrate the model geography as well as other secondary parameters. Fourth, conditional on all other parameters, I internally calibrate association bargaining power and the driver wage. As part of this final step, I also invert the model to obtain home location-specific amenities and work location-specific wages. My model replicates non-targeted moments, including the key Market Mohring and Bus Supply Insurance Effects, and accurately predicts the real-life reported commute modes of stated preference survey respondents.

Finally, I employ the estimated model to demonstrate that cities can attain sizable welfare gains through low-cost policies that alleviate frictions, such as wait time and safety, in their existing privatized shared transit networks. I show that a minibus formalization program of government-set fares and subsidies to associations, broadly analogous to minibus concessions tested in Cape Town and implemented in other cities such as Dakar, can approximate the socially-optimal passenger wait times. The combination of these two instruments simultaneously scales up demand and bus supply to optimize the Market Mohring and Bus Supply Insurance Effects. Wait times fall, particularly on long suburban routes where buses previously filled slowly. Thanks to shorter waits and lower fares, commuters can reallocate towards higher-wage work locations, and modal shifts from cars to minibuses ameliorate road congestion and emissions. Formalization continues to generate similar gains under four model extensions: nested logit commuter demand, lower “non-rush-hour” commuter inflows, endogenous bus departure timing, and agglomeration spillovers to wages. Motivated by specific commuter concerns, I next simulate enforcement-related policies. Minibus speed limit enforcement, which approximates a scale-up of Cape Town’s *Blue Dot Taxi* minibus safety pilot, and government-provided station security guards to ward off robberies and assaults each turn out to increase welfare by over 2 percent.

As a basis for comparison, I also evaluate three alternative policies representative of other types of transport reforms commonly discussed in developing-country cities. A minibus schedule, as attempted in Dakar, among other cities, allows passengers to avoid wait times. In contrast to the formalization program, fares remain high, which limits the increases in minibus use and the ensuing welfare gains. Next, Cape Town’s existing government-provided formal bus rapid transit line, due to low population density, reaches only a limited number of commuters. In consequence, the *MyCiti* line’s welfare benefits, even accounting for commuters’ re-optimization of their home and work locations, fail to outweigh the high construction costs. Finally, were Cape Town to increase minibus size, perhaps through a second round of an existing *Minibus Taxi Recapitalization* program, wait times would increase without substantial offsetting decreases in road congestion. I deliberately exclude from my analysis a series of policies, such as minibus routes with numerous



intermediate stops or optimal networks of bus rapid transit lines, which Cape Town’s highway-based road network and fiscal constraints, respectively, render infeasible. Nonetheless, numerous avenues emerge through which policymakers in rapidly growing yet resource-poor African cities could leverage existing privatized transit networks to better connect the urban fringe and help their residents reach higher-paying jobs.

## Related Literature

The literature has made much progress in understanding the effects of urban *public* transport infrastructure in developing countries. These papers typically leverage newly opened government-run bus rapid transit or metro lines as natural experiments to estimate key parameters of what are known as quantitative spatial models (Ahlfeldt et al. (2015)) and go on to highlight effects on the spatial sorting of workers (Tsivanidis (2023)), gentrification (Balboni et al. (2020)), labor informality (Zarate (2024)), or crime (Khanna et al. (2024)). Kreindler et al. (2023) go a step further and characterize optimal public transit networks. These models excel at predicting how commuters change their home and work locations in response to changes in commute costs, which typically take an exogenous, “iceberg” form. The fixed commute costs in this class of models, however, cannot be applied to study a mode such as privatized shared transit, where wait times, travel times, and fares depend crucially on the endogenous responses of both passengers and transport providers. I contribute the first model of this privatized transit sector and, in contrast to the literature, can thus study low-cost policies which directly leverage the private-sector response.

More recent work uses natural experiments rather than structural estimation to answer related but distinct questions. Björkegren et al. (2025) quantify the response of privatized shared transit to a natural experiment involving the roll-out of Lagos’s new public bus network and derive model-based sufficient statistics for welfare. Mbonu and Eaglin (2024) identify how shorter-run minibuses supply responses to demand shocks and bus breakdowns vary with market structure. More generally, I relate to a broader literature which applies structural models to study key urban questions in developing countries (Bryan and Morten (2019), Kreindler (2022), Hsiao (2023), Gechter and Tsivanidis (2023), Vitali (2024), Harari and Wong (2024), and Hsiao (2024)).

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 institutional context. In Section III, I discuss a series of facts to rationalize my focus on the wait time and safety frictions. 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 minibuses market in Cape Town. Online Appendix B provides additional details regarding all datasets employed in the paper.

## Newly-Collected and Household Survey Data

No existing data could characterize minibus passenger wait times or the prevalence of specific quality-related minibus attributes, such as driving safety or security guards. I fill both gaps 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. Over the course of the 6-10am morning peak commute, 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 impute passengers' wait times, which I set aside to later compare to my model's predictions. 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. The routes in my sample, as I detail in Online Appendix [B.1.1](#), broadly mirror the universe of approximately 429 unique routes within Cape Town in terms of bus traffic, length, and fares paid. I pair these observations of the loading process with data on fares, passenger boardings, and travel times collected onboard randomly-sampled minibus trips on these same routes.

Second, I quantified commuter demand via a stated preference survey composed of a series of discrete choice experiments designed to estimate commuters' values of time and quality as well as price sensitivity. Relative to revealed preference approaches, stated preferences have two key advantages in my setting. First, they permit the measurement of valuations for currently rare or non-existent attributes, such as minibus station security guards. Second, stated preference surveys introduce exogenous variation in these same attributes, which would typically vary systematically with unobserved characteristics of different commute alternatives.

In my survey, I asked respondents to consider a hypothetical work commute trip and then choose a preferred option in a series of *choice sets* composed of two minibus alternatives, as in Figure [1](#). Options varied in five attributes: cost, travel time, and three binary "quality improvements," namely, security guards at the publicly-owned, shared minibus stations, driver adherence to speed limits, and whether the minibus loads more passengers than seats. I restrict attributes to take on realistic values and choose the combinations of attribute values which characterize each alternative to maximize parameter precision via a *d-efficiency* algorithm now standard in the literature (Rose and Bliemer ([2009](#))). Respondents then completed one of two randomized questionnaires of five choice sets each.

The literature has identified a number of key challenges specific to stated preference approaches, including comprehension and *hypothetical bias*, each of which I address ex-ante through survey design and ex-post through robustness tests. When it comes to comprehension, commute mode choice arguably corresponds closely to the kind of "choice[s] among a small number of realistic, familiar, and fully described products" most likely to facilitate respondent understanding (Ben-Akiva et al. ([2019](#)) p. 13-14). Nonetheless, to test the cognitive load on respondents, I conducted a pilot survey, after which I reduced the number of choice sets and alternatives per set. In line with best practices in developing-country contexts (Mangham et al. ([2009](#))), I chose relevant attributes based on past surveys, included pictograms, and ensured that trained enumerators

**FIGURE 1. MINIBUS STATED PREFERENCE SURVEY: CHOICE SET**

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

*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.

guided respondents question-by-question.<sup>3</sup>

*Hypothetical bias*, in turn, refers to distortions related to the fact that respondents do not need to satisfy an actual time or monetary budget constraint. Direct questions regarding willingness to pay particularly invite such distortions, particularly if respondents hope to influence the implementation of particular policies (Whittington 2010). In contrast, my discrete choice experiments, which involve alternatives which vary in multiple dimensions, would require an unusual degree of sophistication to “game” in this fashion. Nonetheless, to test for any such bias, I follow state-of-the-art guidance (Whittington (2010) and Ben-Akiva et al. (2019)) and collect respondents’ *revealed* preferences, i.e. actual commute modes, for later comparison to their stated preferences.

To conduct interviews, enumerators randomly approached  $N = 526$  respondents at one intermodal transport hub and two minibus stations on weekdays. This sampling strategy, 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, so I employ this data only to estimate relative preferences for different *minibus* attributes. Furthermore, I probe the potential for selective non-response based on unobservables, such as individuals’ value of time, by re-estimating the model separately for each type of survey site.<sup>4</sup>

<sup>3</sup>In particular, to choose relevant attributes, I considered the concerns reported by commuters in the 2013 Cape Town Household Travel Survey, displayed in Online Appendix Figure A.3.

<sup>4</sup>The in-person, in-public survey mode, chosen to ensure that respondents carefully completed and comprehended the survey, precluded reliable calculation of response rates and, naturally, the collection of data on non-respondents who refused to speak to enumerators.

I pair my stated preference survey with household survey data collected by the City of Cape Town. The 2013 Cape Town Household Travel Survey provides home and workplace locations as well as incomes for a representative citywide sample of  $N = 22,332$  households. A smaller subsample completed a commute stated preference survey, which involved hypothetical choices between different modes of transport, namely car, formal public transit, or minibus. These options varied in commute time and cost, but not, as in my survey, quality-related attributes. This city-run stated preference effort likewise relied on low-dimensional, familiar choice sets, pictograms, and in-person interviews to boost respondent comprehension and achieve an overall response rate around 90%.

## 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 who use each mode of transport. A full 28% of non-university-educated, henceforth “low-skill,” and 7% of high-skill workers commute via minibuses. Another third of low-skill commuters use limited publicly-provided “formal” transit alternatives, which 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 overwhelming majority of high-skill commuters instead drive to work.

The minibus market consists of a large number of small private firms that own, on average, less than two buses each (Woolf and Joubert (2013)) but collude via minibus *associations*. In particular, city government typically grants a single association per origin-destination pair, or *route*, the exclusive rights to regulate firm entry and fares (Kerr (2018)). Firms pay an entry fee to such an association to operate on a single route. Associations’ chosen fares require city government approval, with a particular eye towards the proportion of “monthly [passenger] income spent on public transport” (City of Cape Town (2014) p. 65).<sup>5</sup> Though the institutional details naturally vary by country, a similar mix of sophisticated collective organization and government regulation defines minibus markets across the African continent (Kerzhner (2022, 2023)).

Minibuses in Cape Town, as in cities from Djibouti to Nairobi to Lilongwe (Kerzhner (2022)), depart at random, unscheduled times from large minibus stations.<sup>6</sup> Stations feature clearly marked loading lanes for each route, where passengers wait in sometimes-sizable queues to board minibuses, as pictured in Figure 3a. Figure 3b depicts the typical case where one minibus per route loads passengers from the front of the queue. In my data, 96% of these minibuses follow the *fill-and-go* practice near-universal among privatized shared transit providers worldwide (Cervero and Golub (2007)) and depart with a full load of at least 15 passengers.<sup>7</sup> After one bus departs, limited bus supply and congestion in and around stations often mean

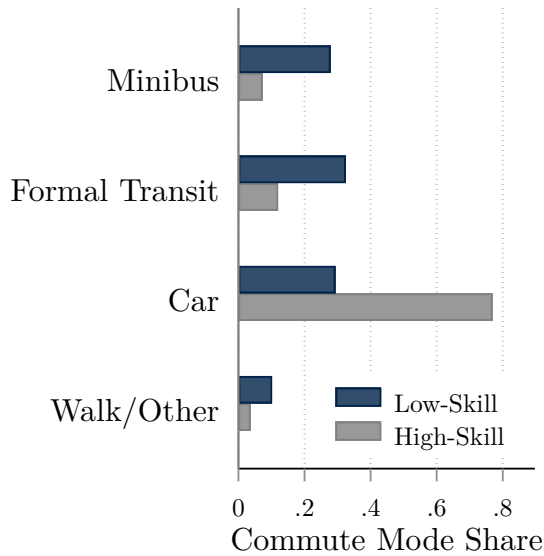
<sup>5</sup>Virtually all firms join associations (Antrobus and Kerr (2019)). Additionally, the law requires individual firms to obtain an effectively pro-forma government permit and operate a vehicle with one of several approved seat capacities (Jobanputra (2018) p. 290). However, up to half of firms lack these permits (City of Cape Town (2014) p. 77).

<sup>6</sup>In Online Appendix Figures A.2a-A.2b, I investigate the minutes of the hour at which minibuses on two representative routes depart and find no systematic patterns that would evidence even an informal, word-of-mouth schedule.

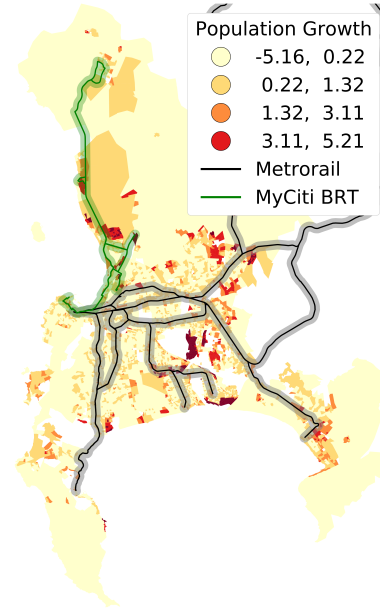
<sup>7</sup>Online Appendix Figures A.1a-A.1b display the distributions of queueing time and the number of buses loading simultaneously in my station count data; during 70% of the peak commute hours I observe, exactly one minibus is loading passengers. Though the

**FIGURE 2. DIFFERENT MODES OF TRANSPORTATION IN CAPE TOWN**

(A) COMMUTE MODE SHARES BY SKILL GROUP



(B) FORMAL PUBLIC TRANSIT NETWORK



*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.

that a gap of several minutes passes before another bus arrives to load passengers. During the rush hours I study, minibuses typically provide a direct origin-destination service, often on highways with few logical intermediate stops. In fact, in my on-board tracking data, 80% of passengers board at the origin and 67% leave the bus at one of the last three stops, close to the destination.

### Existing Reform Efforts

Transport policy in Cape Town has focused largely on the build-out of the aforementioned *MyCiti* bus rapid transit (BRT) network, centered around the single trunk route from the central business district (CBD) to the city’s northern suburbs depicted in Figure 2b. In part due to the long commute distances and low population densities typical of South African cities, *MyCiti*’s ridership and fare revenue have greatly underperformed expectations (Behrens and Farro (2016)).

The resulting fiscal burden on local government has motivated a renewed focus on policies which target the existing minibus sector. Most prominently, a nationwide *Taxi Recapitalization Program* offers minibus owners who agree to scrap an old bus a grant worth 25% of the cost of a new minibus (Schalekamp and McLachlan (2016) and Schalekamp and Klopp (2018)); partly thanks to this effort, older bus models have

law allows for several discrete bus sizes, the 15-passenger-plus-driver variant accounts for 94% of licensed minibuses (Jobanputra (2018) p. 290).

**FIGURE 3. MINIBUS LOADING PROCESS AT ORIGIN STATION**

(A) PASSENGER QUEUE



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



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

almost disappeared. A pilot program known as *Blue Dot Taxi* installed GPS trackers in minibuses and paid incentives to drivers for safe driving behavior; despite wide acclaim from policymakers and passengers, the program has since ended due to funding shortfalls (Ribbonaar et al. (2023)). Finally, in 2019, Cape Town piloted a minibus formalization program known as the *Taxi Operating Company Model* on one route. City officials worked with the route association to implement a fixed schedule, streamline minibus supply and fares, and centralize revenue collection and driver payment (Arroyo-Arroyo, Chevre, Schalekamp, et al. (2021) and Springler et al. (2023)). Though the program did not involve direct subsidies, such payments have been variously discussed in the media and government reports.<sup>8</sup> Indeed, though any reform which endangers minibus associations' rents naturally harbors the potential for political conflict, these various existing efforts set a precedent for successful association-government collaboration.

### III. FACTS ABOUT THE MINIBUS MARKET

Privatized shared transit markets in cities across the developing world suffer from different mixes of frictions; I now present four facts to motivate my focus on wait times and safety in the context of Cape Town.

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

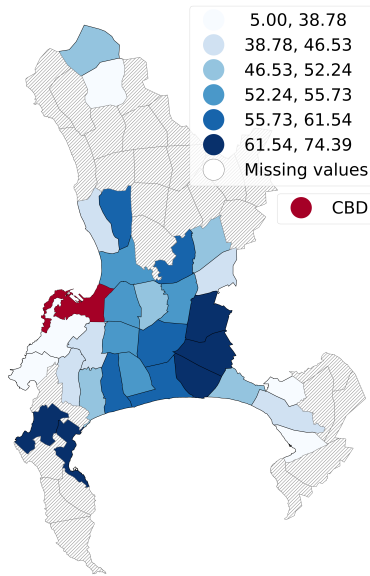
Across the 44 routes in my station counts, passenger wait times—*queueing time* spent in the queue plus the *loading time* it takes the bus to fill up and depart—average 15.4 minutes. Because they frequently take highways and make few intermediate stops, minibuses achieve an end-to-end average speed of 36 kilometers

<sup>8</sup>See, for example, the op-ed “Subsidies for South Africa’s minibus taxis must prioritise needs of passengers – and cities” or a report commissioned by the Competition Commission South Africa.

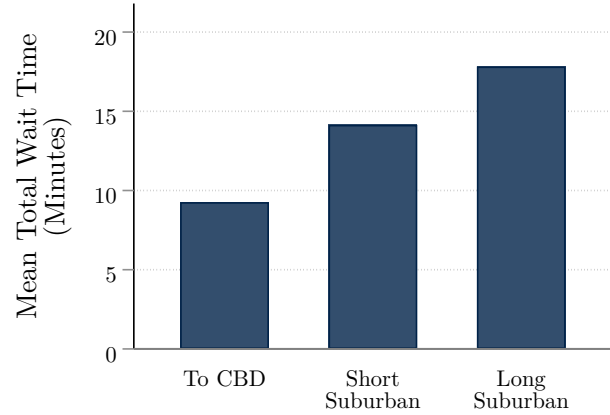


**FIGURE 4. MINIBUS PASSENGERS' TOTAL COMMUTE AND WAIT TIMES**

(A) TOTAL COMMUTE TIMES BY NEIGHBORHOOD



(B) WAIT TIMES BY ROUTE TYPE



*Notes:* Panel (A) maps the mean total commute time by home neighborhood (mesozone) reported by minibus commuters in the 2013 Cape Town Household Travel Survey. Panel (B) displays the mean wait time (mean passenger queueing time plus mean loading time for the bus to fill up) in my station count data on routes of a given type. *To CBD* indicates routes which terminate in the transport analysis zone containing the CBD, while *Short* and *Long Suburban* are those that do not terminate in the CBD, with average distance driven below and above the median, 13.5km, across suburban routes.

per hour, such that wait time makes up over a third of the 43-minute average total minibus commute, i.e. wait plus travel time.<sup>9</sup> Minibus passengers in outlying suburbs, mapped in Figure 4a, face among the longest total commute times, in significant part due to waiting. Indeed, long-distance suburb-to-suburb routes' wait times, as evident in Figure 4b, average almost 20 minutes, almost twice as long as on routes to the CBD.

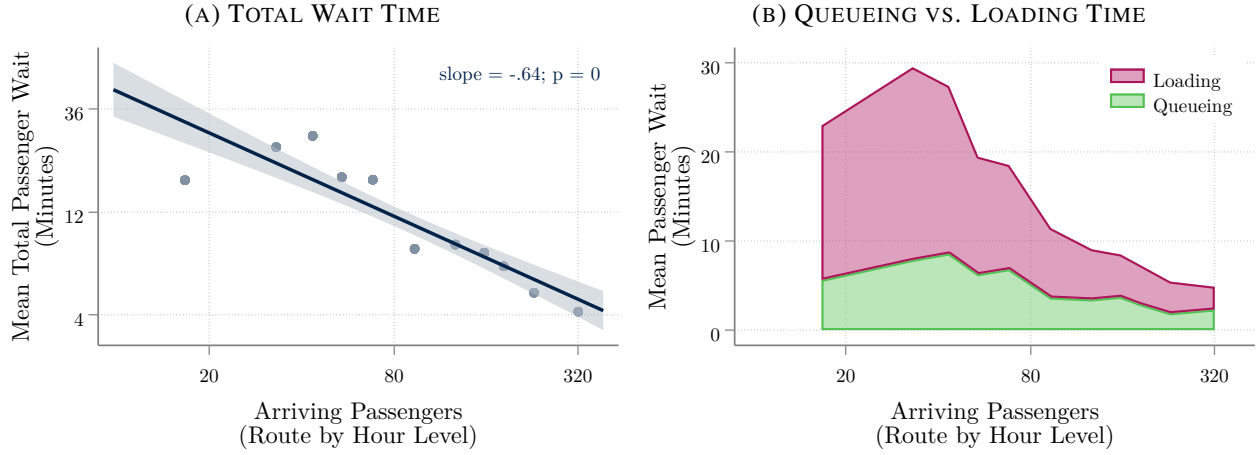
Typically lower-demand suburban routes experience the longest waits precisely because minibuses, like all forms of public transit, benefit from a form of increasing returns to scale via wait times. Indeed, in the case of *government*-provided, formal public transit, Mohring (1972) first observed that higher demand justifies higher frequencies and thus lower wait times. On large bus or metro routes with, for all intents and purposes, unlimited capacity per vehicle, the “Mohring Effect” reflects policymakers’ choices to have vehicles depart more often. In the case of minibuses, their much more binding capacity constraint bakes the Mohring Effect into the technology itself: routes with twice as many passengers require twice as many buses, yielding twice the frequency and half the wait time, on average.

My station count data confirm these increasing returns in wait times, an effect which I henceforth term the *Market Mohring Effect* to reflect its market-based origins. In particular,

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

<sup>9</sup>In comparison, buses in New York City recently averaged 13km/h.

**FIGURE 5. MARKET MOHRING EFFECT**



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

Figure 5a plots the relationship between, on the horizontal axis, the rate of newly-arriving passengers for a given route and hour and, on the vertical axis, mean total wait times. Total waits fall with passenger arrivals because associations must add buses to accommodate the additional passengers, which decreases queueing time, and these buses then fill up in a shorter loading time, as evident in Figure 5b. The Market Mohring elasticity of waits to passengers, at  $-0.64$ , comes remarkably close to the formal transit equivalent of  $-0.67$  estimated by Parry and Small (2009), despite their fundamentally different sources.<sup>10</sup>

The small size of minibuses not only bakes in the Market Mohring Effect but also creates the potential for queues. Indeed, over the course of the day, the total number of seats on minibuses—if they depart full—must equal the number of passengers, but spikes in passenger arrivals over the course of the day may not all fit on the next minibus. In consequence,

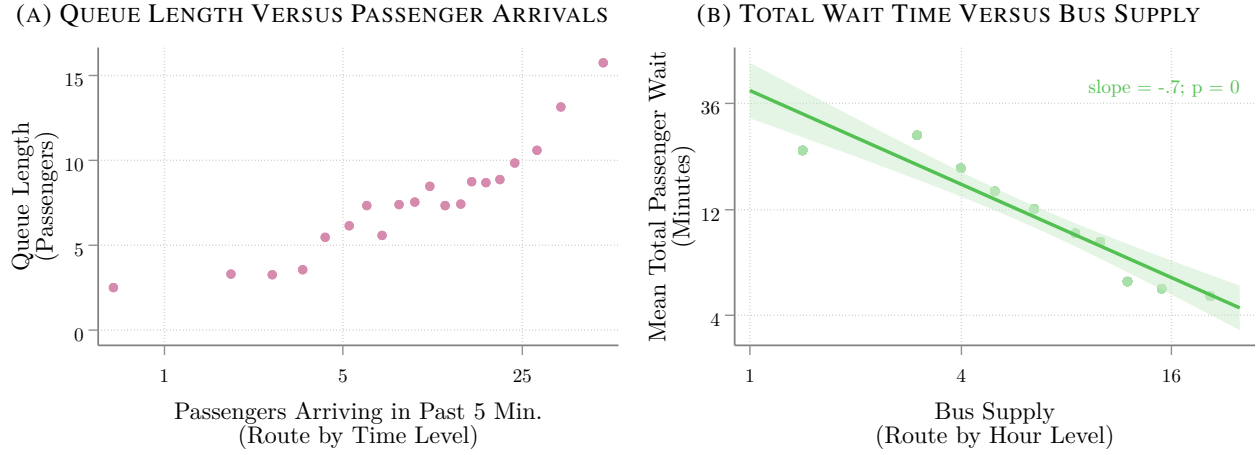
**Fact 3.** *Passenger arrivals lead to the build-up of queues, but additional bus supply acts as “insurance” and decreases queues.*

Figure 6a plots the number of passengers arriving to a route’s loading area within a given five-minute block on the horizontal axis and the resulting number of passengers queueing at the end of the five minutes on the vertical axis. Sudden demand spikes evidently increase queue length in a convex fashion. In Figure 6b, I again plot total passenger wait times, now versus the number of buses operating on a route during a given hour. Intuitively, on routes where the association provides more “surplus” bus supply, buses arrive more frequently, which decreases the likelihood that sudden demand shocks translate into long queues. In turn, as

<sup>10</sup>The comparable elasticity in Parry and Small (2009) equals  $-0.67$ , which I obtain by multiplying their elasticity of wait times with respect to vehicle frequency, at headways below 15 minutes, of  $-1$ , with the calibrated fraction of additional public transit demand accommodated through additional vehicles,  $0.67$ .



**FIGURE 6. BUS SUPPLY AS “INSURANCE” AGAINST QUEUES**



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

observed, total passenger wait times decrease. I henceforth refer to this role of surplus bus supply, whereby extra buses help weather demand shocks and prevent the build-up of queues, as the *Bus Supply Insurance Effect*. In sum, then, policies which aim to decrease suburban minibuses’ high wait times could either increase the scale of passenger demand or induce additional surplus bus supply “insurance.”

Not only wait time but also quality-related frictions ostensibly warrant attention:

**Fact 4.** *Among all attributes of minibuses, commuters express the highest levels of dissatisfaction with road safety and security from crime.*

Indeed, in a 2013 survey, road safety and crime led a list of minibus-related grievances that also included crowdedness, cleanliness, timetable adherence, ease of use, and distance to the stop, as displayed in Online Appendix Figure A.3. More generally, the fundamental public-good nature of security in shared public spaces, such as Cape Town’s largely unguarded minibus stations, rationalizes state intervention.<sup>11</sup> I thus include speed limit enforcement vis-a-vis minibuses and minibus station security guards as attributes in my stated preference survey and conduct associated policy counterfactuals.

## IV. A THEORY OF PRIVATIZED SHARED TRANSIT

In this section, I build a model of the privatized shared transit sector. Passengers wait for minibuses to depart, in a queue and then on board the bus, which gives rise to increasing returns. Minibus associations Nash-bargain fares with the city government and then choose the amount of “reserve” bus supply. On the

<sup>11</sup>Note that only a few select stations currently have security guards; anecdotally, guards could help address commuter concerns regarding the risk of theft or assault.

demand side, 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 constrained monopolist. Minibuses on a route, with common capacity  $\bar{\eta}$ , load commuters from a passenger queue, drive to the route's destination on roads subject to congestion, and then immediately return to the origin to take on more passengers. The single minibus association for each route, with bargaining power  $\beta$ , Nash-bargains fares with the city government and then freely chooses the supply of minibuses. 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  $\theta_i^g$ , a work location, indexed by  $j$  and with fixed wage  $\omega_j^g$ , and one mode  $m \in \{\text{minibus } (M), \text{formal transit } (F), \text{car } (A)\}$  for a single commute to work.

Minibus and car travel times depend on congestion in the road network, composed of all locations  $\mathcal{L}$  as nodes and links between each location  $i$  and some fixed set of neighbors  $S(i)$ . I leave formal transit wait times, travel times, and fares, all of which typically adjust only infrequently and in response to complex political considerations, as exogenous.

## Minibus Loading Process

In my model, passengers wait to board minibuses in a route-specific queue at the origin station. After completing a previous trip, buses arrive to load passengers and only depart once full; together, these elements give rise to passenger wait times.

To model the passenger queue, I build on a fundamental queueing framework in operations research known as the *M/M/1 queue*. In particular, passengers on origin-destination route  $ij$  arrive to the route's first-in-first-out queue at a Poisson rate  $\lambda_{ij}$  determined by the number of commuters who choose to commute from  $i$  to  $j$  by minibus. Whenever a bus is present in the loading area, passengers from the front of the queue board the bus at Poisson rate  $\mu$ . This *queueing efficiency*, inversely proportional to the average time each passenger takes to board, reflects station infrastructure but also behavioral and cultural factors. I then follow Avi-Itzhak and Naor (1963) and augment this otherwise standard M/M/1 queue with so-called “service interruptions” to capture the fact that passengers can only board – i.e. the queue only “operates” – when a bus is present. In particular, in my framework, after a bus departs and leaves the queueing area empty, another arrives at a Poisson rate  $\lambda_{ij}^B$ , upon which passengers may continue to board. Each bus departs at another Poisson rate  $\mu_{ij}^B$ . I next discuss the determination of bus arrival and departure rates.

I impose a functional form for the bus arrival rate  $\lambda_{ij}^B$  consistent with a stylized model of bus operations. In

particular, suppose that the association on a route supplies  $b_{ij}$  minibuses and that these buses, on average, spend  $L_{ij}$  minutes at the station loading passengers, a quantity I henceforth term *loading time*. Furthermore, denote the on-road travel time from  $i$  to  $j$  by  $T_{ij}$ , such that a round trip takes  $2T_{ij}$ . If buses return to the origin empty after dropping off their passengers at the destination and experience no additional congestion in and around the minibus station, the time interval between equally-spaced arrivals of  $b_{ij}$  buses equals  $(L_{ij} + 2T_{ij}) / b_{ij}$ . Since  $L_{ij}$  minutes elapse between a bus's arrival and departure, the gap between one bus's *departure* and the arrival of the next will, for sufficiently small  $L_{ij}$  relative to  $b_{ij}$ , approximately equal  $\frac{2T_{ij}}{b_{ij}} - L_{ij}$ .<sup>12</sup> In reality, due to congestion in and around minibus stations, the next bus may not immediately arrive to the loading area, even on routes with large numbers of excess buses. To capture such in-station congestion, I assume a bus arrival rate  $\lambda_{ij}^B$  such that the expected gap between a bus departure and the arrival of the next satisfies

$$\frac{1}{\lambda_{ij}^B} \equiv \frac{1}{\rho} \log \left\{ \exp \left[ \rho \left( \frac{2T_{ij}}{b_{ij}} - L_{ij} \right) \right] + 1 \right\}. \quad (1)$$

The parameter  $\rho < \infty$ , which I term *arrival efficiency*, controls how quickly this bus gap asymptotes to zero, i.e. how quickly the next bus can take a departed bus's place, as bus supply  $b_{ij}$  increases. Intuitively, greater values of  $\rho$  correspond to milder vehicular congestion within and around minibus stations.

I impose that the bus departure rate  $\mu_{ij}^B$ , in turn, adjusts such that all buses depart with a full load of passengers in expectation, consistent with near-universal practice in Cape Town.<sup>13</sup> If passengers were to continuously board buses, a Poisson departure rate of  $\mu / \bar{\eta}$  would ensure that buses depart with an average of  $\bar{\eta}$  passengers on board. However, under Poisson passenger arrivals, the queue will sometimes shrink to zero, at which point the bus no longer continues to fill. Thus, the bus departure rate consistent with full bus departures must account for the endogenous probability, denoted by  $p_{ij}^{q|b}$ , that a nonzero queue of passengers is waiting and able to board, conditional on a bus being present in the loading area. In particular, I assume that  $\mu_{ij}^B \equiv \mu p_{ij}^{q|b} / \bar{\eta}$ .

I have now built up the framework required to characterize minibus passengers' total wait times. Minibus passengers first wait in the queue; suppose that a passenger arrives to find  $n$  passengers in the queue for a given route and no bus present in the loading lane. Then, the expected *queuing time* until the passenger finishes boarding the bus, denoted by  $Q_{ij}$ , satisfies

$$E(Q_{ij} | n, \text{no bus present}) \equiv \frac{1}{\lambda_{ij}^B} + \frac{n+1}{\mu p_{ij}^b}. \quad (2)$$

The first term in (2) captures the wait for a bus to arrive. The second term equals the expected boarding time per passenger,  $1/\mu$ , times the number of passengers ahead of and including the newly-arrived passenger,

<sup>12</sup>Intuitively, one full "cycle" on the route takes  $L_{ij} + 2T_{ij}$  minutes. To ensure equal intervals between bus arrivals, the next bus must arrive at the origin  $\frac{L_{ij} + 2T_{ij}}{b_{ij}}$  minutes after the first bus arrives and begins loading. At that moment,  $\frac{L_{ij} + 2T_{ij}}{b_{ij}} - L_{ij}$  minutes will have elapsed since the first bus finished loading and departed. For simplicity and motivated by the relative magnitudes in my data, I assume that  $L_{ij}$  is small relative to  $b_{ij}$ , such that  $\frac{L_{ij}}{b_{ij}} \approx 0$  and the bus gap approximately equals  $\frac{2T_{ij}}{b_{ij}} - L_{ij}$ .

<sup>13</sup>In 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 my primary formalization counterfactual.

$n + 1$ , divided by the unconditional probability  $p_{ij}^b$  that a bus is present in the loading area.<sup>14</sup> This probability that a bus is present will logically depend directly on the expected loading time  $L_{ij}$ , relative to the gap  $1/\lambda_{ij}^B$  when no bus loads, such that

$$p_{ij}^b \equiv \frac{L_{ij}}{\frac{1}{\lambda_{ij}^B} + L_{ij}}. \quad (3)$$

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

Together, these expressions highlight how additional bus supply,  $b_{ij}$ , ameliorates queueing times. Extra bus supply, from (1), decreases the gaps between buses, more so, the higher bus arrival efficiency  $\rho$ . From (3), lower bus gaps go hand-in-hand with higher probabilities  $p_{ij}^b$  that a bus is present in the station at all. The greater this probability, the less acutely sudden spikes in the number  $n$  of passengers waiting, in (2), push up queueing times  $Q_{ij}$ . I thus refer to this benefit of additional buses as the *Bus Supply Insurance Effect*, the insurance being against demand shocks. The benefit of such insurance increases with both bus arrival efficiency  $\rho$  and passenger queueing efficiency  $\mu$ . To understand the latter case, consider a low  $\mu$ : when passengers board slowly and  $L_{ij}$  rises, the next bus, again from Equation (1), will more likely already have shown up by the time the previous has filled.<sup>15</sup>

Once a passenger boards, he or she waits an additional loading time which, given my assumptions on the bus departure rate  $\mu_{ij}^B$  and the usual Poisson properties, equals

$$L_{ij} \equiv \frac{1}{\mu_{ij}^B} = \frac{\bar{\eta}}{\mu p_{ij}^{q|b}} \quad (4)$$

in expectation. Loading time, in turn, directly incorporates the *Market Mohring Effect* discussed in Section III: higher demand  $\lambda_{ij}$  increases the probability of a nonzero queue  $p_{ij}^{q|b}$ , so buses fill faster, and passenger wait times fall. The larger vehicles themselves, i.e. the higher  $\bar{\eta}$ , the greater the scope for these increasing returns.

In sum, passenger demand  $\lambda_{ij}$  and the bus supply  $b_{ij}$  jointly determine the total time passengers wait in the queue and for the bus to fill. In Appendix A.1, I detail how I follow Brancaccio et al. (2024) and simulate the evolution of the queue *out* of steady-state. As a result, I obtain expected queueing time  $Q_{ij}$  and loading time  $L_{ij}$  for each route.

## Associations

Associations, which each possess exclusive rights to a single minibuss route, make two key decisions. First, a route's association Nash-bargains fares with the city government, and second, conditional on fares, the

<sup>14</sup>Intuitively, imagine a bus is present only half the time; in that case, the time elapsed, even after an initial bus arrives, until the newly-arrived passenger boards will be twice as long as if a bus were continuously present and able to board passengers.

<sup>15</sup>To build intuition, consider the scenario with more than  $\bar{\eta}$  passengers queueing; then,  $L_{ij} = \bar{\eta}/\mu$  until the queue clears, so lower queueing efficiency directly increases loading times, and Equation (1) will imply a low bus gap even at low levels of bus supply.

association chooses bus supply to maximize its flow of profits. I discuss and solve the association problem via backward induction.

### Bus Supply

A minibuss association's bus supply,  $b_{ij}$ , continuously runs trips on its designated route. Individual minibuses arrive to the origin minibuss station, load passengers for, on average,  $L_{ij}$  minutes, drive to the route's destination in a congestion-affected travel time  $T_{ij}$ , and return to the origin.

Associations earn revenue from fares, denoted by  $\tau_{ij}$ , and pay operations costs and driver wages. Each trip on a route costs  $\chi_{ij}$  in operations costs; drivers earn a fixed wage  $\bar{\omega}$  per unit time. The flow of minibuss association profits,  $\Pi_{ij}$ , thus equals the arrival rate of busloads times, in brackets, profits per trip, net of wages paid to the drivers of the  $b_{ij}$  total buses in circulation at any instant:

$$\Pi_{ij} \equiv \underbrace{\frac{\lambda_{ij}}{\eta}}_{\text{Busloads}} \left[ \underbrace{\bar{\eta} \tau_{ij}}_{\text{Revenue}} - \underbrace{\chi_{ij}}_{\text{Ops. costs}} \right] - \underbrace{\bar{\omega} b_{ij}}_{\text{Driver wages}} \quad (5)$$

Associations then choose bus supply  $b_{ij}$  to maximize profits, conditional on bargained fares; the associated first-order condition implies that

$$\left( \tau_{ij} - \frac{\chi_{ij}}{\eta} \right) \frac{\partial \lambda_{ij}}{\partial b_{ij}} = \bar{\omega}. \quad (6)$$

Recall that additional bus supply lowers passenger wait times via the Bus Supply Insurance Effect and thus increases demand. The left-hand side of (6) captures the resulting marginal revenue, which the association sets equal to the marginal cost of bus supply, the driver wage  $\bar{\omega}$ . The former marginal revenue expression directly highlights how higher fares  $\tau_{ij}$  motivate the association to internalize a greater share of the value of “extra” bus supply.

### Fares

Before, but in anticipation of this optimal supply choice, the association on each route separately Nash-bargains fares with the 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 Yürükoğlu (2022)), here replicates Cape Town's government-mediated fare scheme. Associations, with bargaining power  $\beta$ , seek to maximize profits,  $\Pi_{ij}$ . In line with the official route approval guidance discussed in Section II, the city government's objective equals a straightforward measure of commuter ability to pay, namely average destination workplace income  $\omega_j$  minus the minibuss fare.<sup>16</sup> I assume that both sides receive zero payoff if bargaining fails, so fares,  $\tau_{ij}$ , maximize  $\Pi_{ij}^\beta (\omega_j - \tau_{ij})^{1-\beta}$ . The

<sup>16</sup>Note that, for simplicity, this average workplace income uses weights equal to a skill group's share of the aggregate city population, so that  $\omega_j \equiv \left( \sum_{g'} N^{g'} \right)^{-1} \sum_g N^g \omega_j^g$ .

resulting fares,

$$\tau_{ij} = \omega_j - \frac{1 - \beta}{\beta} \frac{\Pi_{ij}}{\frac{\partial \Pi_{ij}}{\partial \tau_{ij}}}, \quad (7)$$

rise with association bargaining power; as  $\beta$  approaches unity, fares reach the pure monopoly level which satisfies  $\frac{\partial \Pi_{ij}}{\partial \tau_{ij}} = 0$ . The higher fares, of course, the lower passenger demand – with pernicious effects on total wait times, given the Market-Mohring increasing returns.

## 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  $\theta_i^g$ ; (ii) the deterministic commute value  $U_{ijm}^g$ ; (iii) the work location's fixed wage  $\omega_j^g$ ; and (iv) a Gumbel-distributed idiosyncratic preference  $\varepsilon_{ijm}$  with variance scaled by the parameter  $v$ .<sup>17</sup>

The commute value  $U_{ijm}^g$  depends on mode-specific utility costs, commute time, and monetary costs and linearly approximates a micro-founded commute model, detailed in Online Appendix C.1. In particular, minibus commuters receive commute value

$$U_{ijM}^g \equiv -\underset{\substack{\uparrow \\ \text{utility cost}}}{\kappa_M^g} - r\omega_j^g \left( \underset{\substack{\uparrow \\ \text{queueing time}}}{Q_{ij}} + \underset{\substack{\uparrow \\ \text{loading time}}}{L_{ij}} + \underset{\substack{\uparrow \\ \text{travel time}}}{T_{ij}} \right) - \underset{\substack{\uparrow \\ \text{fare}}}{\tau_{ij}}. \quad (8)$$

The skill-specific utility cost  $\kappa_m^g$  reflects commuters' non-pecuniary taste for a particular mode, which might depend on quality-related factors such as crime risk or driving safety. In turn, the *product* of the time preference rate  $r$  and the wage  $\omega_j^g$  determines commuters' disutility of the total commute time in parentheses. Finally, commuters' sensitivity to the fare,  $\tau_{ij}$ , varies inversely with the Gumbel scale parameter  $v$ . Commuters on other modes make similar tradeoffs. Formal transit commuters pay a utility cost  $\kappa_F^g$  as well as exogenous fares  $\tau_{ijF}$ , wait a fixed headway  $H_{ij}$ , and travel a fixed  $T_{ijF}$  minutes to their destination, so that  $U_{ijF}^g \equiv -\kappa_F^g - r\omega_j^g (H_{ij} + T_{ijF}) - \tau_{ijF}$ .<sup>18</sup> Car commuters similarly pay a utility cost  $\kappa_A^g$  as well as a fixed monetary cost  $\tau_A$  and spend the same on-road, congestion-affected travel time  $T_{ij}$  as minibuses; they thus receive 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,

<sup>17</sup>I 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, and 3.1% of minibus riders report also using a car or motorcycle.

<sup>18</sup>Note that, as detailed in Online Appendix D.4.5, I allow formal commuters to walk to transit stops and transfer between buses and trains in my calibration of  $H_{ij}$  and  $T_{ijF}$ , such that a formal transit route exists between any origin and destination.

whereby skill- $g$  commuters choose to commute by mode  $m$  from home  $i$  to work  $j$  with a probability

$$\pi_{ijm}^g = \frac{\exp\left(\theta_i^g + U_{ijm}^g + \omega_j^g\right)^{1/\nu}}{\sum_{i',j',m'} \exp\left[\theta_{i'}^g + U_{i'j'm'}^g + \omega_{j'}^g\right]^{1/\nu}}. \quad (9)$$

Minibus demand then translates into passenger arrivals,

$$\lambda_{ij} \equiv \sum_g N^g \pi_{ijM}^g, \quad (10)$$

such that local amenities and wages determine, via commuter location choices, the extent to which a minibus route grows busy and leverages the Market Mohring Effect.

## Road Congestion

I further allow for road congestion: minibuses and cars increase each others' travel times through the road network, as follows. For tractability, I require that vehicles follow the sequence of network links,  $s(i, j)$ , that minimizes free-flow travel time from  $i$  to  $j$ . Vehicle inflow  $x_{kk'}$  onto an individual network link equals the sum of the commuter inflow, adjusted by vehicle capacity, over origins and destinations whose path crosses that link, i.e.  $x_{kk'} = \sum_{i',j':kk' \in s(i',j')} \sum_{g'} N^{g'} \left( \pi_{i'j'M}^{g'}/\bar{\eta} + \pi_{i'j'A}^{g'} \right)$ . The travel time over an individual link in the road network then depends on a link-specific free-flow travel time  $\bar{t}_{kk'}$  and vehicle inflow  $x_{kk'}$  according to  $t_{kk'} \equiv \bar{t}_{kk'} x_{kk'}^\gamma$ , where  $\gamma$  denotes the road congestion elasticity. Finally, travel time for minibuses and cars equals the sum of travel times over road network links along the corresponding path,  $T_{ij} \equiv \sum_{kk' \in s(i,j)} t_{kk'}$ . Importantly, due to their higher capacity, minibuses impose a lower per-commuter congestion externality, so shifts towards minibuses might actually *reduce* road congestion, to the extent that commuters switch out of cars.

## Equilibrium and Efficiency

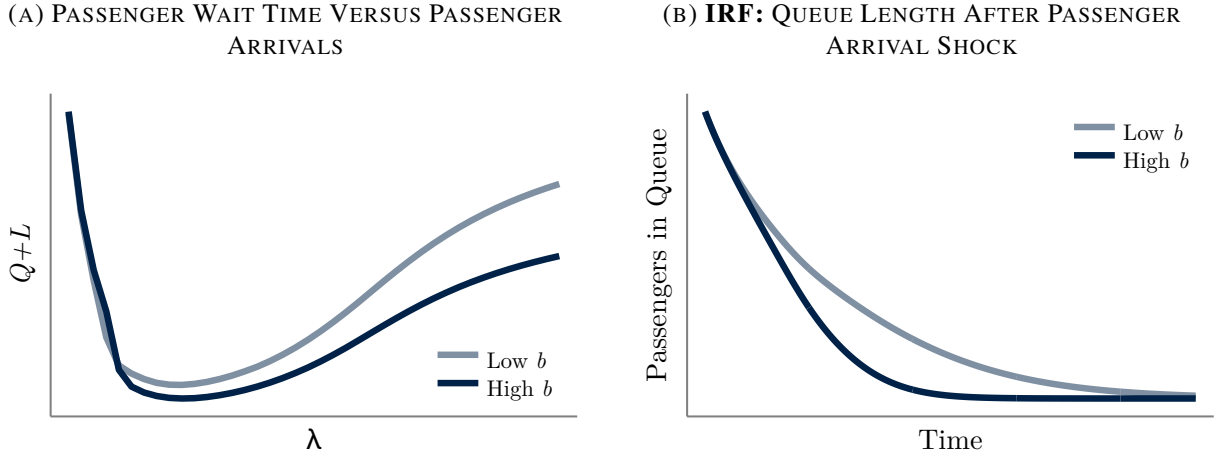
Equilibrium in my model depends on a set of parameters and an exogenous model geography which determine minibus supply, minibus fares, and commuter choice probabilities, as follows.

**Definition. (Equilibrium)** *Given parameters  $\{r, \nu, \kappa, \tau_A, N, \mu, \bar{\eta}, \rho, \bar{\omega}, \beta, \gamma\}$  and the model geography  $\{\theta, \omega, \chi, \bar{t}, H, T_F, \tau_F\}$ , an equilibrium is a vector  $\{b, \tau, \pi\}$  such that (i) associations maximize profits according to (6); (ii) the Nash bargaining condition (7) holds; and (iii) commuter demand satisfies (9).*

### Market Mohring and Bus Supply Insurance Effects

To build intuition, I now graphically demonstrate how passenger wait times vary with demand and bus supply. In Figure 7a, I simulate total passenger wait time  $Q + L$  for different rates of passenger arrivals  $\lambda$ , separately for lower and higher levels of bus supply  $b$ . As demand increases from low levels, buses fill faster and wait times sharply drop, regardless of the extent of extra bus supply “insurance.” These increasing returns in wait time correspond exactly to the Market Mohring Effect; thanks to minibuses' fundamental capacity constraint, routes with more passengers must have more frequent departures, on average. As demand increases further,

**FIGURE 7. MARKET MOHRING AND BUS SUPPLY INSURANCE EFFECTS**



*Notes:* Panel (A) plots simulated queueing plus loading time for one route in my data and with parameters as later estimated, versus the passenger arrival rate  $\lambda$ , for a low and high value of bus supply  $b$ . Panel (B) displays the impulse-response function for the number of passengers in the queue, again for parameters as estimated, in response to a one-time simultaneous arrival of 30 passengers. Note that, to calculate  $\lambda^B$  in the latter simulations, I hold loading time fixed.

however, the large clusters of passenger arrivals that lead to the build-up of queues occur more frequently, so average queueing times slowly start to rise.

Wait times rise more slowly, however, in the high-bus-supply case: surplus buses allow for continuous loading and thus “insure” against demand shocks. This Bus Supply Insurance Effect represents a new margin unique to privatized shared transit and its typically smaller vehicles. To further highlight how minibus supply dampens queues, in Figure 7b, I plot the impulse-response function of queue length, in number of passengers, to a (one-time) passenger arrival shock. On a formal transit route with large vehicles, the resulting “queue” would disappear upon the arrival of the next metro or bus. In the case of minibuses, in contrast, the queue clears rapidly only with sufficient surplus bus supply such that multiple buses arrive in quick succession.

### *Efficiency*

I now define welfare and, as a benchmark, the social planner problem. I denote commuters’ ex-ante expected Gumbel utility as  $\bar{\Omega}^g \equiv v \log \left[ \sum_{i',j',m'} \exp \left( \theta_{i'}^g + U_{i'j'm'}^g + \omega_{j'}^g \right)^{1/v} \right]$ . Then, consistent with an economy where commuters receive rebates of their share of aggregate minibus profits upon birth,

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

$$\Omega \equiv \sum_g N^g \bar{\Omega}^g + \Pi. \quad (11)$$

The social planner then chooses allocations subject to the queueing and road congestion technologies. I



represent the numerically-simulated queueing and loading times which define the former as  $Q(\lambda_{ij}, b_{ij})$  and  $L(\lambda_{ij}, b_{ij})$ , respectively.

**Definition.** (*Planning Problem*) The social planner chooses bus supply  $b$  for each route as well as skill-group-specific commute choice probabilities  $\pi$  for each home, work, and mode to maximize welfare,  $\Omega$ , as in (11), subject to the queueing technology,  $Q_{ij} = Q(\lambda_{ij}, b_{ij})$  and  $L_{ij} = L(\lambda_{ij}, b_{ij})$ , the road congestion technology  $T_{ij} = \sum_{kk' \in s(i,j)} \bar{t}_{kk'} x_{kk'}^\gamma$ , as well as the constraint  $\sum_{i,j,m} \pi_{ijm}^g = 1$  for each skill group  $g$ .

Crucially, my model features three potential sources of inefficiency: (i) the Market Mohring Effect of demand on minibus passenger wait times; (ii) the classic Road Congestion Externality imposed by minibus and car commuters; and (iii) the Bus Supply Insurance Effect of extra buses on passenger wait times. I now formally define each. Note that both the Market Mohring Effect and the Road Congestion Externality operate through commuter demand. To conceptually separate the two, I define each while holding either the congestion-affected on-road travel times or the Market-Mohring-affected minibus queueing and loading times fixed, respectively.

**Definition. (Market Mohring Effect)** Given an equilibrium  $\{b, \tau, \pi\}$ , associations internalize the Market Mohring Effect whenever demand  $\pi$  maximizes welfare,  $\Omega(b, \cdot)$ , conditional on bus supply  $b_{ij}$  and on-road travel times  $T_{ij}$  and subject to  $Q_{ij} = Q(\lambda_{ij}, b_{ij})$ ,  $L_{ij} = L(\lambda_{ij}, b_{ij})$ , the road congestion technology  $T_{ij} = \sum_{kk' \in s(i,j)} \bar{t}_{kk'} x_{kk'}^\gamma$ , and the adding-up constraint.

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

**Definition. (Bus Supply Insurance Effect)** Given an equilibrium  $\{b, \tau, \pi\}$ , associations internalize the Bus Supply Insurance Effect whenever  $b$  maximizes welfare,  $\Omega(\cdot, \pi)$ , conditional on commuter choice probabilities and subject to  $Q_{ij} = Q(\lambda_{ij}, b_{ij})$  and  $L_{ij} = L(\lambda_{ij}, b_{ij})$ .

However, in the status quo, where Nash bargaining over fares gives associations a form of *limited* market power, they generically neither induce the Market Mohring-efficient demand level nor provide the optimal amount of bus supply insurance.

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

$$\tau_{ij} = \frac{\chi_{ij}}{\bar{\eta}} + \sum_{g'} N^{g'} \pi_{ijM}^{g'} r \omega_j^{g'} \left( \frac{\partial Q_{ij}}{\partial \lambda_{ij}} + \frac{\partial L_{ij}}{\partial \lambda_{ij}} \right) \quad (12)$$

and the Bus Supply Insurance Effect if and only if

$$\tau_{ij} = \frac{\chi_{ij}}{\bar{\eta}} - \left( \frac{\partial \lambda_{ij}}{\partial b_{ij}} \right)^{-1} \sum_{g'} N^{g'} \pi_{ijM}^{g'} r \omega_j^{g'} \left( \frac{\partial Q_{ij}}{\partial b_{ij}} + \frac{\partial L_{ij}}{\partial b_{ij}} \right). \quad (13)$$

Commuters fail to internalize the Road Congestion Externality.

*Proof.* See Appendix A.2. □

In other words, associations induce optimal wait times only in a double-knife-edge, Hosios (1990)-type case, when association bargaining power  $\beta$  happens to coincide with the fares which fully price in (i) the effect of demand on wait times on the right-hand side of (12) and (ii) the value of additional bus supply “insurance” on the right-hand side of (13).

In all other cases, the limited form of market power introduced by Nash bargaining allows the association to set fares too high for passenger demand to reach the optimal scale and fully leverage the Market-Mohring increasing returns in wait time. Simultaneously, however, the fact that fares generally fall below monopoly levels means that associations do not fully internalize the insurance against queue build-up which extra bus supply provides.<sup>19</sup> Naturally, absent appropriate road pricing, car and minibus commuters will not account for the road congestion spillovers they impose.<sup>20</sup> Thus, the market power common to many privatized shared transit markets, as parameterized by  $\beta$ , has both benefits and costs, in efficiency terms. On the one hand, market power facilitates the provision of Bus Supply Insurance in the form of higher  $b_{ij}$ . On the other hand, greater  $\beta$  induces higher fares which restrict minibus demand far below its efficient scale, given the Market-Mohring increasing returns.

## Discussion

My model abstracts away from a number of dimensions for reasons of data availability and tractability. First, I do not model production and thus cannot endogenize wages or allow for agglomeration economies. My setting lacks the data on commercial floorspace use and wages by skill group necessary to calibrate a firm production function, as in Tsivanidis (2023). Moreover, Cape Town lacks recent infrastructure investments which substantially shifted home and work locations and could thus provide variation to quantify agglomeration spillovers. I note also that the literature generally estimates only moderate agglomeration elasticities (Combes and Gobillon (2015)). Nonetheless, in Online Appendix E.3, I extend the model to include agglomeration spillovers and simulate my primary formalization counterfactual using an agglomeration elasticity from the literature.

Second, I do not allow for heterogeneous home and work location choice elasticities. I could tractably do so in a nested logit framework but similarly lack the necessary exogenous variation to quantify these elasticities in the South African context. To again explore the robustness of my central findings, I solve a version of my model with nested logit demand in Online Appendix E.3.

Third, my model focuses exclusively on rush-hour work commutes. Resource constraints prevented the collection of station count data throughout the entire day, but my estimated queueing and bus arrival frameworks could be combined with one of several state-of-the-art models of non-work trips (Miyauchi

---

<sup>19</sup>Note that, though higher  $\beta$  typically induces associations to choose closer-to-efficient bus supply  $b_{ij}$ , even in the case of  $\beta \rightarrow 1$ , associations will not set exactly efficient bus supply when there are multiple worker types due to a Spence (1975)-type quality distortion.

<sup>20</sup>Note that association bargaining power could in-theory induce minibus fares which exactly happen to price in road congestion, but, in my model, the same will never be true for car commuters.

et al. (2022)). A suggestive simulation of lower commuter demand in Online Appendix E.3 suggests that, outside of peak commute hours, more routes would likely resemble my current low-demand routes, which particularly benefit from the realization of Market-Mohring increasing returns.

Fourth, I do not allow for associations or minibus drivers to have their buses depart less-than-full, in line with the fact that virtually all buses in my data do in fact depart full. However, in Online Appendix E.3, I consider a tractable extension, whereby associations optimally set the bus departure rate  $\mu^B$  and thus the average number of passengers carried. An additional extension would allow drivers to pick up passengers in the origin neighborhood en route, at the cost of longer travel time.

Fifth, my model does not explicitly include all margins of interaction between different associations and between associations and government. Note that the association choice of bus supply does feature strategic interactions and essentially corresponds to a Cournot equilibrium. The Nash bargaining of fares captures one association-government interaction and similarly occurs in a Nash-equilibrium sense, conditional on all other such bargains. However, I could extend my model to include other, extra-legal forms of interaction between associations or allow associations to bargain with government over margins other than fares.

Finally, the increasing returns associated with the Market Mohring and the Bus Supply Insurance Effects introduce the potential for equilibrium multiplicity. I note, however, that road congestion as well as the congestion inherent in my bus arrival rate function could, at sufficiently large values of passenger arrivals and bus supply, counterbalance the increasing returns and thus induce a unique equilibrium. Though I cannot rule out the existence of additional equilibria analytically, I initialize my solution algorithm at  $\lambda_{ij}$  and  $b_{ij}$  half and twice as high as in my baseline computed equilibrium; in each case, the solution algorithm converges back to virtually the same equilibrium.<sup>21</sup>

## V. ESTIMATION OF MINIBUS AND DEMAND PARAMETERS

I estimate the model’s structural parameters in four main steps. First, I devise instrumental variables strategies which employ the station counts of passengers and buses to identify the queueing efficiency  $\mu$  and arrival efficiency  $\rho$ . Second, from commuters’ stated preferences, I estimate their mode-specific utility costs  $\kappa_m^g$ , rate of time preference  $r$ , and Gumbel shock scale  $v$ . Third, I estimate the road congestion elasticity from TomTom traffic data and externally calibrate a geography composed of  $I = 18$  *transport analysis zones* in Cape Town, as well as other secondary parameters. Fourth, conditional on all other parameters, I internally calibrate the association bargaining power  $\beta$  and driver wage  $\bar{w}$ . As part of this final step, I also invert the model to obtain home location-specific amenities  $\theta_i^g$  and work location-specific wages  $\omega_j^g$ . Table 1 summarizes all calibrated parameters.

---

<sup>21</sup>In particular, in no case does the median absolute deviation between different starting points of an equilibrium object exceed  $2.3 \times 10^5$ , where I take median absolute deviations across routes.

**TABLE 1. CALIBRATED PARAMETERS**

Parameter	Description	Value	Parameter	Description	Value
<i>Externally Calibrated</i>			<i>Stated Preference</i>		
$I$	Number Locations	18	$r$	Commuter Rate of Time Pref.	0.001
$N^g$	Commuter Populations		$v$	Gumbel Shape	4.76
$\bar{t}_{ij}$	Free-Flow Driving Time		$\kappa_M^l$	Low-Skill Minibus Util. Cost	7.7
$H_{ij}$	Formal Wait Time		$\kappa_M^h$	High-Skill Minibus Util. Cost	15
$T_{ijF}$	Formal Travel Time		$\kappa_F^l$	Low-Skill Formal Util. Cost	3.6
$\tau_A$	Car Commute Cost	5.2	$\kappa_F^h$	High-Skill Formal Util. Cost	9.2
$\tau_{ijF}$	Formal Fare		<i>Internally Calibrated</i>		
$\bar{\eta}$	Minibus Capacity	15	$\beta$	Assoc. Bargaining Power	0.1
$\chi_{ij}$	Minibus Operating Cost		$\bar{\omega}$	Driver Wage	0.005
<i>Minibus Loading</i>			<i>Model Inversion</i>		
$\mu$	Queueing Efficiency	4.56	$\theta_i^g$	Amenities	
$\rho$	Arrival Efficiency	0.18	$\omega_j^g$	Wages	
<i>Road Congestion</i>					
$\gamma$	Road Congestion Elasticity	0.09			

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

### *Minibus Loading*

To quantify the ability of minibus routes to absorb additional passengers, I require estimates of two key parameters of the minibus loading process: the passenger queueing efficiency  $\mu$  and bus arrival efficiency  $\rho$ . I quantify queueing efficiency, which controls the extent to which passenger arrivals generate queues, based on the relationship between demand and observed bus loading times. In a similar vein, I estimate the bus arrival efficiency, which determines how readily additional buses mitigate queues, from the effect of bus supply on the time intervals between successive buses. To permit a model-based discussion of identification, I lay out an empirical version of my queueing framework in Online Appendix C.2 that incorporates (i) unobserved loading delays, which end at Poisson rate  $\iota_{ij}$ , and (ii) an unobserved bus arrival speed, denoted by  $\varsigma_{ij}$ .

First, I devise a new strategy to estimate queueing efficiency,  $\mu$ . Existing methods, typically applied to virtual computer queues, would require observation of the exact time individual passengers take to board buses (Asanjarani et al. (2021)) or very-high-frequency observations of queue length (Chowdhury and Mukherjee (2011, 2013)). Were there always a queue of passengers and no interruptions to loading, the properties of the resulting standard M/M/1 queue would imply that I could estimate queueing efficiency directly from the average loading time divided by the number of passengers. However, this boarding process may periodically pause when the queue empties of passengers or when other unobserved interruptions, such as weather, impede

loading. In either case, the total loading time is no longer directly proportional to  $\mu$ . I instead estimate queueing efficiency from a model-consistent relationship that accounts for these pauses in loading.

In particular, I leverage the extent to which the prevalence of passenger queues translates into lower bus loading times. Equation (4), translated to the route  $ij$  by hour  $h$  by date  $d$  level, implies that expected bus loading times depend on the probability  $p_{ijhd}^{q|b}$  of the existence of a (nonzero) queue of passengers, conditional on a bus being present, or “queue prevalence” for short. I calculate the latter in my station counts as the share of every-5-minute observations where I see one or more passengers in the queue *and* a bus present in the loading area, divided by the total share with a bus present. I then regress the loading time  $L_{sijhd}$  of bus  $s$  route  $ij$  departing during hour  $h$  on date  $d$  on hourly queue prevalence  $p_{ijhd}^{q|b}$  according to

$$L_{sijhd} = \frac{\bar{\eta}}{\mu} \left( p_{ijhd}^{q|b} \right)^{-1} + \underbrace{\delta_{ijd} + \delta_{ihd} + \zeta_{sijhd}}_{\equiv \varepsilon_{sijhd}}. \quad (14)$$

I parameterize the unobservable term  $\varepsilon_{sijhd}$ , which contains the unobserved rate  $\iota_{sijhd}$  at which loading delays end, as the sum of route-by-date and origin-by-hour-by-date fixed effects as well as an idiosyncratic shock.<sup>22</sup> Intuitively, the greater queueing efficiency, the more queue prevalence  $p_{ijhd}^{q|b}$  translates into lower loading times, since any passengers in the queue will board faster.

The aforementioned interruptions of the usual loading process, for reasons other than the lack of a queue, pose the primary threat to identification. The direction of bias is ex-ante ambiguous: disruptions which coincide with higher demand and thus more frequent queues would bias  $\mu$  towards zero, but shocks such as rain, which might directly depress passenger arrivals, could bias  $\mu$  upwards. Route-by-date fixed effects,  $\delta_{ijd}$ , capture time-invariant causes of interruptions, such as cramped stations, and origin-by-hour-by-date fixed effects,  $\delta_{ihd}$ , absorb any higher delay-susceptibility of the busiest commute times. Thus, the error term  $\zeta_{sijhd}$  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 queue prevalence: the number of commuters from  $i$  to  $j$  who leave for work during hour  $h$ , measured in a 2013 survey.<sup>23</sup> Intuitively, higher commuter numbers should translate directly into the kind of spikes in demand which increase the prevalence of queues.

This distribution of work start times instrument satisfies the exclusion restriction, provided it does not vary systematically with idiosyncratic interruptions to loading. Since I measure queue prevalence and bus loading times in 2022, such 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.1.2, I search for evidence of such systematic hourly trends, which would likely go hand-in-hand with a higher variance of

<sup>22</sup>Recall that I detail how I incorporate the unobservable  $\iota_{sijhd}$  into the baseline model in Online Appendix C.2. Furthermore, note that all routes starting at a given origin  $i$  were surveyed only on one date, so that the  $d$  subscript is redundant, but I nevertheless index by both hour and date for clarity.

<sup>23</sup>At 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) Bus loading time	(2) Bus loading time	(3) Bus loading time
$\bar{\eta}/\mu$ <i>Minibus Capacity/Queueing Efficiency</i>	2.45 (1.12)	2.01 (0.79)	3.29 (1.93)
<i>Anderson-Rubin C.I. for <math>\bar{\eta}/\mu</math>:</i>	[1.21,12.93]	[1.01,5.05]	[1.21, $\infty$ )
Implied $\mu$ <i>Queueing Efficiency</i>	6.12 (2.80)	7.45 (2.93)	4.56 (2.67)
Route-by-Date FE		✓	✓
Origin-by-Hour-by-Date FE			✓
Observations	1101	1101	1101
First-Stage F Statistic	6.16	12.21	3.69

*Notes:* Robust standard errors in parentheses, clustered at the route-by-hour-by-date level. Table presents estimates of (14) over bus departures in my station count data, with fixed effects included, as noted, for route-by-date and origin-by-hour-by-date. I instrument for queue prevalence, i.e. the probability of a nonzero passenger queue conditional on a bus being present in the loading area, using the log number of commuters living in the mesozone spatial unit where the route originates and working in the route's destination mesozone who report leaving their home during hour  $h$ , calculated from the 2013 Cape Town Household Travel Survey. I additionally report the effective first-stage F statistic following Olea and Pflueger (2013) as well as the Anderson-Rubin confidence intervals robust to weak instruments. Finally, I obtain the implied value of  $\mu$  by taking the reciprocal of the coefficient on (inverse) queue prevalence and multiplying by the externally-calibrated minibus capacity  $\bar{\eta}$ .

observable proxies at the corresponding times of day. I examine the variance of wind speeds across days as a proxy for the temperature and rain shocks more likely to affect loading and find no appreciable variation by hour; similarly, the variance of traffic speed across individual vehicles displays no clear hourly trends.

Table 2 displays the queueing efficiency values implied by instrumental variables estimation of (14). I start with a baseline specification in Column 1; in Columns 2 and 3, I add route-by-date and origin-by-hour-by-date fixed effects, respectively. The first stages, presented in Online Appendix Table D.1, in some cases have marginally-low F statistics, so I additionally calculate Anderson-Rubin confidence intervals robust to weak instruments.<sup>24</sup> The efficiency  $\hat{\mu} = 4.56$  in my preferred specification in Column 3 indicates that the average commuter takes about 13 seconds to board a minibus.<sup>25</sup> The literature offers little precedent for comparison, but this short boarding time suggests that most minibus routes could easily absorb additional passenger arrivals without an attendant build-up of queues.

Second, I quantify the bus arrival efficiency  $\rho$ . I estimate an equation for the expectation of the  $gap_{si jhd}$  in minutes between the departure of bus  $s$  from the origin station and the arrival of the next on route  $ij$  during hour  $h$  on date  $d$ . From Equation (1), this conditional expectation decreases with the number of buses  $b_{ijhd}$ ,

<sup>24</sup>The full specification in Column (3), with clustered standard errors, lacks the power to generate a bounded confidence interval.

<sup>25</sup>Note that I obtain the implied value of  $\mu$  by taking the reciprocal of the coefficient on (inverse) queue prevalence in (14) and multiplying by the externally-calibrated minibus capacity  $\bar{\eta}$ .

more so, the greater arrival efficiency:

$$\text{gap}_{sijhd} = \frac{1}{\rho} \log \left\{ \exp \left[ \rho \left( \frac{2T_{ij}}{b_{ijhd}} - L_{sijhd} \right) \right] + 1 \right\} + \underbrace{\Gamma_{n(i)hd} + \Gamma_{n(j)} + \vartheta_{sijhd}}_{\equiv -\frac{1}{\rho} \log \zeta_{sijhd}}. \quad (15)$$

The error term, here, contains the unobserved arrival speed  $\zeta_{sijhd}$ , which, recall, I formally introduce in Online Appendix C.2. Intuitively,  $\zeta_{sijhd}$  reflects factors such as overall foot and car traffic levels in origin neighborhoods and is thus likely correlated with associations' chosen bus supply. I control for such congestion with origin neighborhood  $n(i)$ -by-hour-by-date fixed effects and additionally include destination neighborhood  $n(j)$  fixed effects, where neighborhood refers to a Transport Analysis Zone in Cape Town. Only route-specific idiosyncratic slowdowns in  $\vartheta_{sijhd}$  thus remain to pose a threat to identification. Hence, I estimate (15) via GMM and instrument for the first term on the right-hand side, a function of bus supply, with (log) 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 which originate from the same neighborhood.

In my full preferred GMM specification, I find an arrival efficiency of  $\hat{\rho} = 0.178$ .<sup>26</sup> As a consequence of the implied congestion in bus arrivals, the gap between buses will never fall below around two minutes on the average route in my data. This minimum gap then effectively imposes a limit on the extent to which additional buses can alleviate queues.

### Stated Preferences

Next, I require estimates of key demand parameters that govern commuters' value of time and minibus quality improvements, and thus the scale of welfare gains from improvements therein. Because the stated preference surveys introduce exogenous variation in these commute attributes, I can directly estimate 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  $\xi_z^g$  of each improvement  $z$  on mode utility costs into Equation (9) 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

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

<sup>26</sup>Robust standard error = 0.09,  $N = 1,158$  bus departures. I present full results in Online Appendix D.2 and, consistent with my model, restrict the mean of each set of fixed effects to equal zero.

<sup>27</sup>Unlike in the main text, I include a final higher-order term—the product of fares and wait time—as implied by the micro-founded model in Online Appendix C.1.



Differences in mode shares, all other attributes equal, identify relative utility costs,  $\kappa_m^g$ .<sup>28</sup> The sensitivity of respondents' choices to the presence of quality improvements identifies their effects,  $\xi_z^g$ , on utility costs. Crucially, the extent to which wait as well as travel time,  $T_i$ , differentially decrease the choice probabilities of respondents with higher income  $\omega_i$  helps quantify the rate of time preference,  $r$ . Finally, the variation in fares,  $\tau_i$ , translates into the Gumbel scale  $\nu$  and allows me to calculate dollar willingness to pay for each attribute.

The left panel of Table 3 provides some of the first 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 of the 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)). This substantially lower value, even relative to existing developing-country estimates for India (Kreindler (2022)), might reflect more limited housework and childcare commitments or a lack of available leisure opportunities in residential, often slum, neighborhoods. Next, I estimate a Gumbel scale parameter  $\hat{\nu} = 4.76$ ; this high idiosyncratic preference variance generates 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. The comparatively inelastic demand for minibuses, then, stems from a limited availability of close substitutes as well as, more speculatively, strong personal preferences over modes, perhaps related to the ability to commute with friends or family.<sup>29</sup>

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)). The estimated effects of my three quality improvements, in the right panel of Table 3, help unpack this perhaps atypical distaste for minibuses in South Africa. All three quality improvements – station security guards to deter muggings and harassment, bans on loading more passengers than seats, and enforcement of speed limits – significantly decrease minibuses' utility cost for the low-skill group. Even more strikingly, the high-skill would pay a full \$2.75 per commute for station security. Not surprisingly, in light of South Africa's longstanding violent crime problem, the annual equivalent of this value exceeds women's already sizable willingness to pay for safer walking routes in India in Borker (2021) by up to a factor of five.

To build confidence in the external validity of these demand estimates, I now investigate the influence of selection on both observables and unobservables. Recall that the stated preference samples mirror the population along every observable dimension except, in the case of my own survey, actual commute modes. Thus, I re-estimate the logit model in (16) and weight my survey by realized commute mode choices; the estimates, in Online Appendix D.3, change little. As for unobservables, might it be, for example, that only

<sup>28</sup>While my own survey over-sampled minibus commuters, my survey included only minibus alternatives and so does not contribute to the identification of the relative utility costs across modes.

<sup>29</sup>In contrast, Cape Town drivers, with an implied own-price elasticity of  $-1.62$ , substitute readily towards other modes. Note that, to calculate demand elasticities, I use median minibus fares from my on-board tracking data as well as the median calibrated 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	
$r$	.001		<i>Effects on Utility Costs</i>	<i>Low-Skill</i>	<i>High-Skill</i>
<i>Commuter Rate of Time Pref.</i>	(.0004)		$\xi_{\text{security}}$	-1.09	-2.75
$v$	4.76		<i>Station Security</i>	(0.390)	(0.84)
<i>Gumbel Scale</i>	(1.26)		$\xi_{\text{no overloading}}$	-1.38	-1.39
<i>Utility Costs</i>	<i>Low-Skill</i>	<i>High-Skill</i>	<i>Overloading Ban</i>	(0.437)	(0.543)
$\kappa_F$	3.63	9.17	$\xi_{\text{follows speed limit}}$	-1.36	-0.825
<i>Formal Transit Utility Cost</i>	(0.51)	(1.89)	<i>Speed Limit Enforcement</i>	(0.445)	(0.465)
$\kappa_M$	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.

commuters with lower values of time selected into my survey? The fact that my sample includes representative fractions of observable groups with typically-higher values of time, such as college-educated workers and women (Borghorst et al. (2021)), points towards limited selection along this unobservable dimension.<sup>30</sup> Any remaining selection might logically occur more severely in locations, such as minibus stations, where enumerators intercepted respondents mid-commute. In this vein, I re-estimate the demand model only among respondents interviewed at locations other than minibus stations and, nonetheless, in Online Appendix D.3, obtain similar results.

I finally return to the key challenges in stated preference estimation discussed in Section II, namely comprehension and hypothetical bias. Imperfect comprehension would introduce substantial noise and thus likely attenuation bias. The fact that I obtain large, relatively precise estimates for the values of the three quality-improvement attributes hence provides an albeit imperfect indication that respondents successfully understood the task at hand. In the context of discrete choice experiments with multiple dimensions of costs, e.g. time and monetary, the effects of hypothetical bias on *relative* valuations cannot be conclusively signed. I must instead appeal to the comprehensive over-identification test of respondents' ability to accurately predict their actual behavior presented in Section VI.

#### *External Calibration*

I externally calibrate the road congestion elasticity, secondary parameters, and the model geography. I quantify the road congestion elasticity  $\gamma$  with data from TomTom's Traffic Stats API on traffic volume  $x_{ih}$  and travel time  $t_{ih}$  on all road segments  $i$  across Cape Town during hours  $h$  of a sample day. My model of road congestion implies  $\log t_{ih} = \bar{t}_i + \gamma \log x_{ih} + \varepsilon_{ih}$ , where the unobservable  $\varepsilon_{ih}$  would include disruptions such

<sup>30</sup>See Online Appendix Table B.2. Note also that no respondent in my survey broke off the questionnaire before finishing all discrete choice experiments suggests that the time burden did not significantly impede respondents' commutes.

**TABLE 4. INTERNAL CALIBRATION**

Moment			Parameter		
<i>Description</i>	<i>Data</i>	<i>Model</i>	<i>Description</i>	<i>Value</i>	
Median Minibus Fare (\$)	1.05	1.05	$\beta$ Association Bargaining Power		0.1
Median Queue Length (Number Passengers)	4	4	$\bar{\omega}$ Minibus Driver Wage		0.005

*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.

as weather and special events (Barwick et al. 2022). In Column 1 of Online Appendix Table D.4, I obtain  $\hat{\gamma} = 0.0917$ . Next, I directly obtain the car commute cost  $\tau_A$ , minibus capacity  $\bar{\eta}$ , and the model geography from the data. The latter includes commute populations  $N^s$ , minibus operating costs  $\chi_{ij}$ , free-flow driving times  $\bar{t}_{kk'}$ , and formal transit wait and travel times  $\{H_{ij}, T_{ijF}\}$  as well as fares  $\tau_{ijF}$ .<sup>31</sup> I provide additional details regarding all externally calibrated parameters in Online Appendix D.4.

#### *Internal Calibration and Inversion*

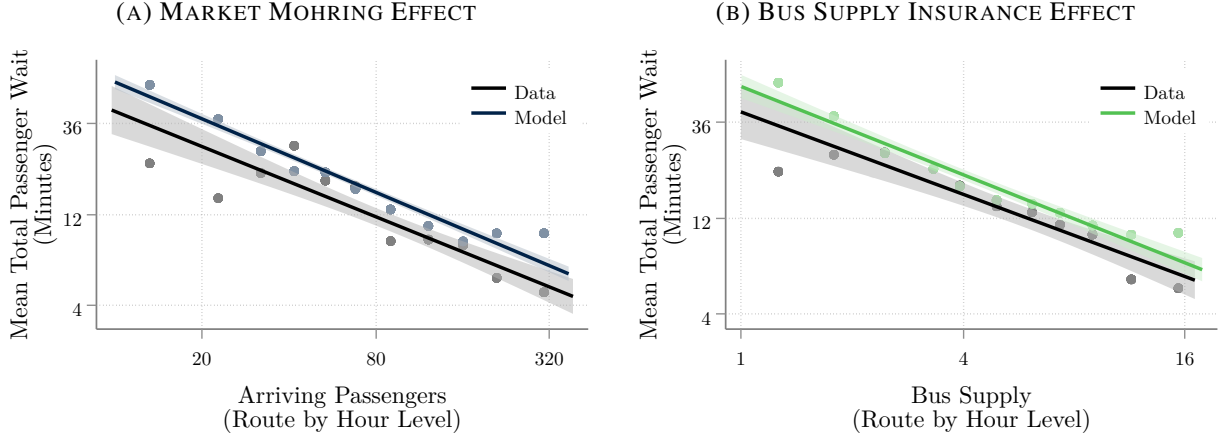
Conditional on all aforementioned parameters, I then match aggregate moments in my newly-collected minibus data to obtain the remaining supply parameters. For each parameter iteration, I invert the model to obtain location-specific amenities and wages. Specifically, I match (i) the median minibus fare,  $\tau_{ij}$ , across routes in the onboard tracking data, and (ii) the median queue length, measured in number of passengers, across routes and five-minute periods in the station counts. As in standard spatial models, I then exactly match each location's observed employment by residence and workplace to calibrate home location amenities  $\theta_i^s$  and workplace wages  $\omega_j^s$ , respectively.<sup>32</sup>

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  $\beta$ , and queue length identifies the minibus driver wage  $\bar{\omega}$ . The final column lists the calibrated parameter values. By way of interpretation, the modest association bargaining power reveals that the city government effectively caps fares far short of monopoly levels, contrary to popular perceptions of association market power. As an initial over-identification test, I verify whether my minibus driver wage matches observed driver earnings. The calibrated value of  $\bar{\omega}$ , in USD per minute, corresponds to approximately 5 ZAR per hour, only slightly below the median hourly earnings of 7.4 ZAR reported by Antrobus and Kerr (2019) for minibus drivers in South Africa.

<sup>31</sup>Since the fixed typically dominate the variable costs of car ownership, I calibrate a single  $\tau_A$ , constant across origins and destinations, based on estimates of the total car ownership cost per (half) day, as discussed in Online Appendix D.4. I calculate formal transit fares  $\tau_{ijF}$  directly from Cape Town's distance-based public transit fare scheme.

<sup>32</sup>I measure employment by residence and workplace in the 2013 Cape Town Household Travel Survey and normalize, for each skill group, (i) the amenity of one location to zero and (ii) the average wage to the empirical skill-group average.

**FIGURE 8. WAIT TIMES IN DATA VERSUS MODEL**



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

## VI. MODEL-PREDICTED MINIBUS OPERATIONS AND AGGREGATE PATTERNS

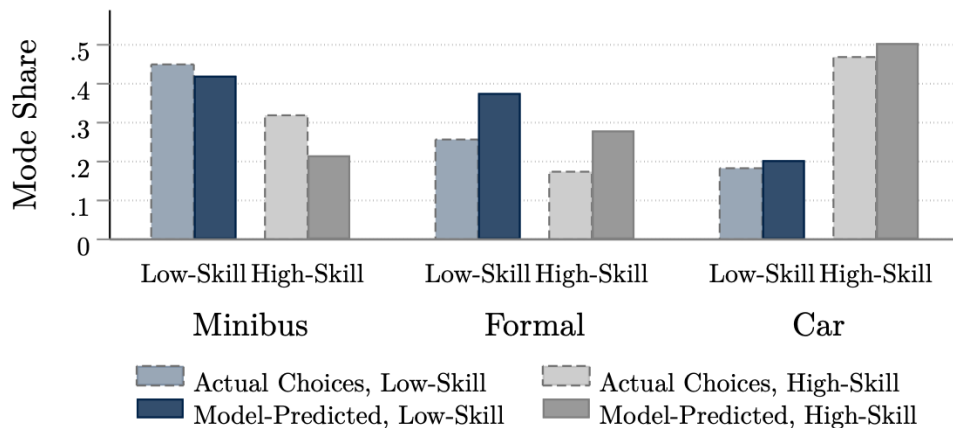
I now demonstrate that my model matches non-targeted data: the minibus network, the Market Mohring and Bus Supply Insurance Effects, and the actual commute modes of stated preference respondents. First, the model-predicted bus supply on each origin-destination pair mirrors the number of distinct minibus routes between each pair of neighborhoods, as mapped in Online Appendix Figures A.4a-A.4b.<sup>33</sup> Importantly, the model matches the concentration of minibuses in central neighborhoods and the direct links between outlying suburbs. To characterize the fit of this model-predicted bus entry, I confirm, in Online Appendix Figure A.4c, that the model reproduces the empirical relationship between passenger arrivals and bus supply.

Second, the model replicates the Market Mohring and Bus Supply Insurance Effects. Figures 8a and 8b again plot total wait times versus passenger arrivals and bus supply, respectively. I show binned scatterplots across route-by-hour observations in the station count data in black and across routes in the model in color. As in the data, model-predicted wait times benefit from Market-Mohring increasing returns and fall with the scale of passenger demand. Similarly, the model successfully predicts the degree to which extra bus supply insures against demand spikes and thus diminishes waits.

Third, I compare stated preference respondents' reported "real-life" commute modes, not employed in demand estimation, to those predicted by my model, estimated based on their stated preferences. This comparison serves as an, albeit indirect, test for the hypothetical bias discussed in Section II. The bars in Figure 9 indicate the skill-group-level shares of respondents in the combined stated preference sample who commute via each

<sup>33</sup>Note that, in reality, unlike in my model, many neighborhood (transport analysis zone) pairs are linked by multiple distinct minibus routes.

**FIGURE 9. 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  $\pi_{ijm}^g$ ; the model predictions are sums of these choice probabilities within each skill and mode.

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 used in estimation quite closely. 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, even though respondents did not suffer the actual monetary or time costs associated with their stated choices, the latter nonetheless seem to accurately predict 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 network and represent broad classes of programs often discussed in relation to privatized shared transit. As a benchmark, I characterize the social planner’s optimum. I then evaluate two sets of policies targeted directly at the frictions highlighted in Section III.

First, I evaluate a minibus “formalization” program of government-set fares and per-passenger subsidies to associations, which approximates the socially-optimal wait times. This policy in many dimensions mimics a largely successful formalization effort in Dakar, which involved concession contracts with groups of minibus owners to operate government-defined routes with regulated fares (Barrett et al. (2016)) at set frequencies. Original plans also included subsidies to cover operating deficits (Filho et al. (2015) and Arroyo-Arroyo, Chevre, Ferro, et al. (2021)). Similar concessions, which variously regulate fares or bus supply and often involve payments from government to minibus operators, exist in cities as diverse as Santiago (Chile), Bogota, Hong Kong, and Quito (Behrens et al. (2016), Rodriguez et al. (2017), and Hurtubia and Leonhardt (2021)). In sub-Saharan Africa, Dar es Salaam has considered similar formalization programs for feeder routes to BRT (Mfinanga and Mafinda (2016)), and Cape Town itself piloted *Taxi Operating Companies*, as detailed in Section II.

Second, I consider the enforcement of minibus speed limits and the provision of security at the publicly-owned minibus stations. The former exercise could approximate the gains from an expansion of Cape Town’s Blue Dot GPS tracker pilot, discussed in Section II, or could leverage digital fleet management systems run by associations, as Nairobi has made steps to introduce (Jennings et al. (2016)). Private security guards, in turn, already watch over most privately-owned public spaces in South Africa, so city government could immediately hire them to guard minibus stations.

For comparison, I then evaluate a series of prototypical policies which Cape Town and cities across sub-Saharan Africa have considered in order to improve or replace privatized shared transit. Cities such as Dakar have attempted to impose schedules on privatized minibus routes (Filho et al. (2015)). Though driver monitoring problems have typically doomed these efforts, the potential wait time gains from a successful implementation motivate my simulation of a schedule for routes in Cape Town with below-median baseline demand. Next, international organizations typically recommend the construction of formal “bus rapid transit” (BRT) lines with dedicated stations and exclusive bus lanes. I thus characterize the net welfare effects of Cape Town’s one existing *MyCiti* BRT line. Finally, cities such as Dakar and Maputo have, through fleet upgrade loans or via outright bans, attempted to encourage the adoption of larger minibuses, ostensibly to alleviate congestion (Romero de Tejada et al. (2023) and Olvera et al. (2024)). I thus consider a second round of Cape Town’s existing Minibus Taxi Recapitalization Program, which, unlike the original, stipulates the purchase of larger 23-passenger minibuses. For simplicity, I simulate a scenario with universal adoption.<sup>34</sup>

My menu of policies deliberately focuses on the most feasible, context-appropriate margins. The existing Minibus Taxi Recapitalization program has essentially eliminated the problem of poor vehicle quality. Furthermore, the nature of Cape Town’s road network, where the most efficient routes between many neighborhoods use limited-access highways, is not well-suited to minibus routes with many intermediate stops. Finally, due to the limited success of the existing BRT line and the constrained local fiscal capacity, I do not consider the build-out of an optimal BRT network.

For each policy, I present changes in the welfare measure  $\Omega$  defined in (11) as equivalent variation: the

---

<sup>34</sup>In addition to the ubiquitous Quantum/HiAce 15-passenger minibuses, Toyota sells 23-passenger *Coasters* in South Africa.

proportionate change in a skill group’s wages  $\omega_j^g$ , at baseline values of  $\{b, \tau, Q, L, T, \kappa, \bar{\eta}, H, T_F\}$ , that leaves the average commuter equally well off as in the counterfactual. I also compute an equivalent variation measure which accounts for the social cost of per-passenger local pollutant and greenhouse gas emissions on each mode, as I further detail in Online Appendix D.4.8.<sup>35</sup> Table 5 at the end of this section summarizes all counterfactuals.

## Social Planner Optimum

As a benchmark, I solve the planning problem defined in Section IV to characterize the socially-optimal commuter choices  $\pi$  and minibus entry  $b$ . I first discuss the optimal commuter choices. The fact that Cape Town has relatively large- $\bar{\eta}$  minibuses generates substantial Market-Mohring increasing returns; intuitively, the larger the buses, the greater the scope for reductions in loading time. In the status quo, associations’ limited form of market power from  $\beta > 0$  keeps demand inefficiently low; in consequence, the planner scales up the number of passengers on all routes, with a median increase of 16%. Demand would particularly increase on routes to higher-wage work locations, as evident in the scatterplot of changes in minibus passenger arrivals  $\lambda_{ij}$  versus average wages at a route’s destination in Figure 10a. High-wage routes previously had higher fares, all else equal, and thus operated farthest below their efficient scale, given the high incomes on offer. The routes with the largest demand increases, mapped in Figure 10c, include both radial trunk routes to the central business district (CBD) as well as routes which connect far-flung suburbs. On all of these routes, association market power previously suppressed demand, such that buses filled inefficiently slowly and further discouraged commuters from accessing the highest wages.

Scaled-up minibus demand, then, requires an accompanying increase in bus supply to ward off the build-up of long queues. The benefit of additional bus supply increases with both passenger queueing efficiency  $\mu$  and bus arrival efficiency  $\rho$ . The higher the former, the more likely passengers will have fully boarded one bus before the next arrives, and the greater the latter, the more substantially additional bus supply will reduce the gaps between them. However, because bargaining restricts fares, associations fail to internalize these benefits. In consequence, as again plotted in Figure 10b versus destination wages, the planner would increase bus supply everywhere – to a greater degree, to match passenger arrivals, on routes to higher-wage destinations. Exactly the newly-viable suburb-to-suburb routes on which the planner scales up demand also, in Figure 10d, require the most additional Bus Supply “Insurance.”

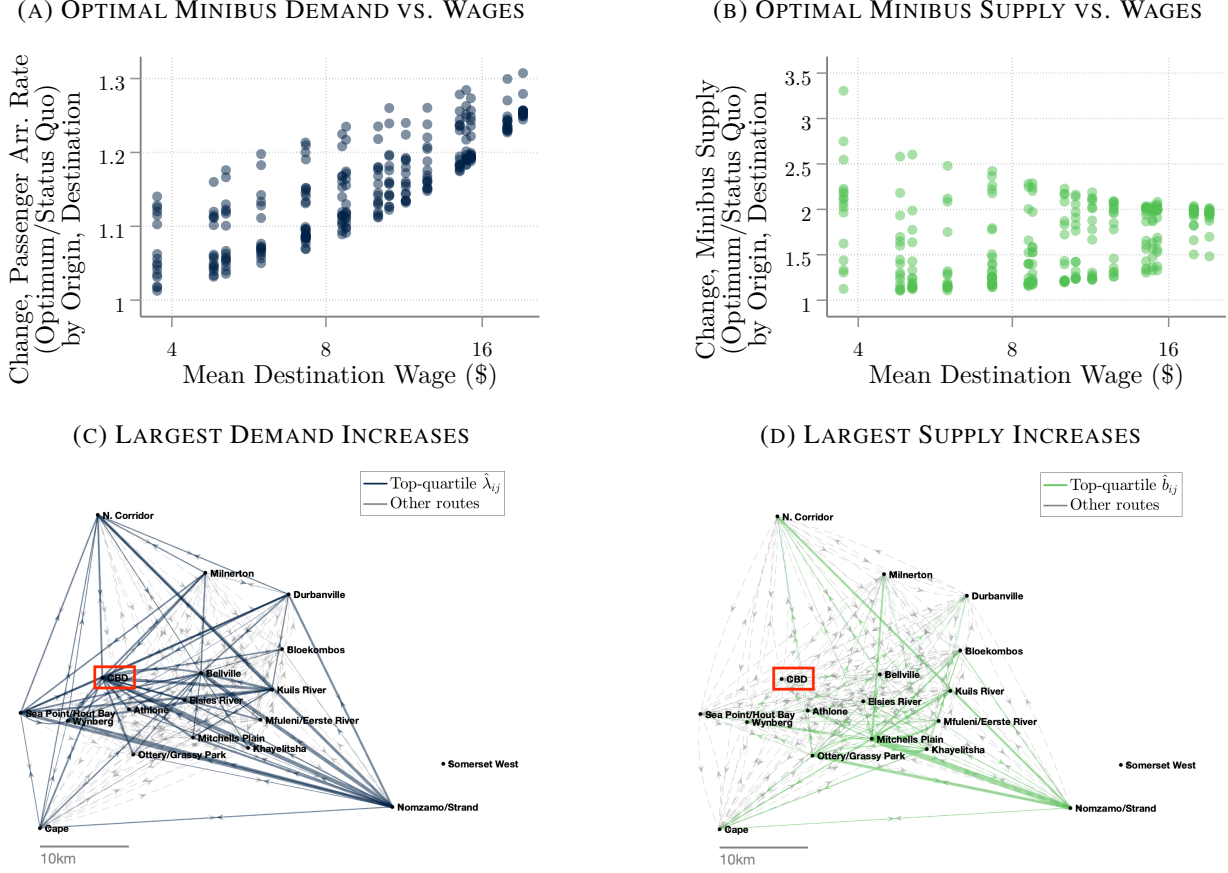
In sum, under the commuter choices and bus supply that optimize wait and travel times, low-skill commuters gain 0.38% in equivalent variation, net of emissions and relative to the status quo, and the high-skill gain 0.17%.<sup>36</sup> Though the raw numbers involved might at first glance appear small, these *net-of-cost* gains rival estimates in the literature for much more substantial urban infrastructure investments. I now compare a series

<sup>35</sup>To calculate welfare at the skill-group level in a manner unaffected by minibus fares, I rebate minibus profits as follows: for each route, I multiply route-level profits by a group’s share among minibus commuters on that origin-destination and then sum across routes (i.e. origin-destination pairs). I allocate emissions costs as well as the monetary costs of the MyCiti and security guard counterfactuals according to population shares.

<sup>36</sup>To calculate skill-group-level welfare, I allocate minibus operations costs and driver wage costs on a given route to each skill group according to their share of minibus commuters on that route, sum across all routes, and subtract the total from that skill group’s total ex-ante expected commuter utility  $N^g \bar{\Omega}^g$ .



**FIGURE 10. SOCIAL PLANNER OPTIMUM**



Notes: Panels (A) and (B) display, on the vertical axis, the social planner's optimal minibus passenger arrival rate  $\lambda_{ij}^*$  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. Panels (C) and (D) map the 25% of routes with the largest proportionate increases in demand,  $\lambda_{ij}$ , and bus supply,  $b_{ij}$ , respectively, from the status quo to the social optimum. Line width indicates the magnitude by which the social optimum exceeds the status quo, and the CBD is outlined in red.

of specific urban transport policies, akin to those already discussed in contexts similar to Cape Town, to this benchmark.

## Minibus Formalization

I begin with a policy tailored to optimize passenger wait times: a minibus formalization program of government-set fares and per-passenger subsidies to associations.<sup>37</sup> In particular, I calculate the minibus fares in (12) which induce associations to internalize Market-Mohring increasing returns. I then find the per-commuter subsidies to associations which close the gap between, on the one hand, these net fares and, on the other hand, the gross fares in (13) which ensure that the association sets efficient bus supply.<sup>38</sup>

<sup>37</sup>Given my focus on frictions which directly impact commuters, my formalization program does not involve other provisions often discussed in the literature on informality, e.g. regulation of labor standards or tax enforcement.

<sup>38</sup>In Appendix A.3, I detail the decentralization of the social optimum. In the formalization counterfactual, commuters of a given skill pay equal lump-sum taxes to fund the association subsidies which correspond to their own minibus use. I calculate the latter by

Optimal fares turn out approximately linear in (log) distance, and subsidies are approximately uniform across routes. The formalization program thus consists of (i) government-set, distance-based minibuses fares  $\bar{\tau}_{ij}$ , equal to a two-piece linear approximation, plotted in Figure 11a, to the optimal fares; and (ii) a uniform per-passenger subsidy  $\bar{z}$  to all associations, equal to the mean of the optimal subsidies. As in Dakar, the Cape Town municipal government would likely implement such a scheme through concession contracts to one association per route. Though any changes to associations' operating rights might prove politically divisive, the formalization scheme I propose simply modifies the existing, already roughly distance-based but informally-negotiated fare scheme. Furthermore, formalization turns out to increase not only commuter welfare but also the after-subsidy profits of associations on every single route.

This approximate optimization of minibuses wait times leverages both Market-Mohring increasing returns and the Bus Supply Insurance Effect to decrease wait times, as plotted in Figure 11b versus route distance. Total queueing plus loading time falls particularly starkly on long routes. These routes previously attracted the lowest passenger numbers and thus suffered the longest wait times; when demand and bus supply rise in tandem, they experience large absolute declines in the queueing time spent waiting for a bus to arrive. Table 5 also highlights the extent of the mode shift from cars to *shared* minibuses, which contributes to less-congested roads and the modest decreases in on-road travel times also evident in Figure 11b.

As minibuses routes reach their optimal scales, inefficiently long waits no longer encumber commuters' choices of home and work locations. Figure 11c plots the changes in work location choice probabilities versus wages and reveals significant spatial reallocation towards the highest-wage locations. The commute flows that grow the most, mapped in Figure 11d, span the aforementioned suburban links and radial trunk routes. These shifts contribute to a 0.45% increase in average earned wages, as Table 5 highlights. Furthermore, as the car mode share decreases, greenhouse-gas and other pollutant emissions fall by over 3%.<sup>39</sup> As a result of these changes in wait time, road congestion, home and work locations, and emissions, the formalization program attains over 90% of the welfare gains available from the full optimization of commuter demand and minibuses supply. Thus, minibuses formalization, which requires no infrastructure and mirrors initiatives implemented across sub-Saharan Africa, has the potential to boost aggregate welfare, dampen emissions, and reduce inequality across skill groups.

#### *Decomposition: Market Mohring vs. Bus Supply Insurance*

Which source of increasing returns drives the benefits of formalization? To tease apart the contributions of the Market Mohring and Bus Supply Insurance Effects, I simulate separate counterfactuals which impose only (i) the formalization fare scheme, with association subsidies solely to close the gap between these lower optimal and the status-quo fares; and (ii) the formalization scheme's per-passenger association subsidies, with status-quo fares charged to passengers. The former "fares only" policy leverages only the Market Mohring

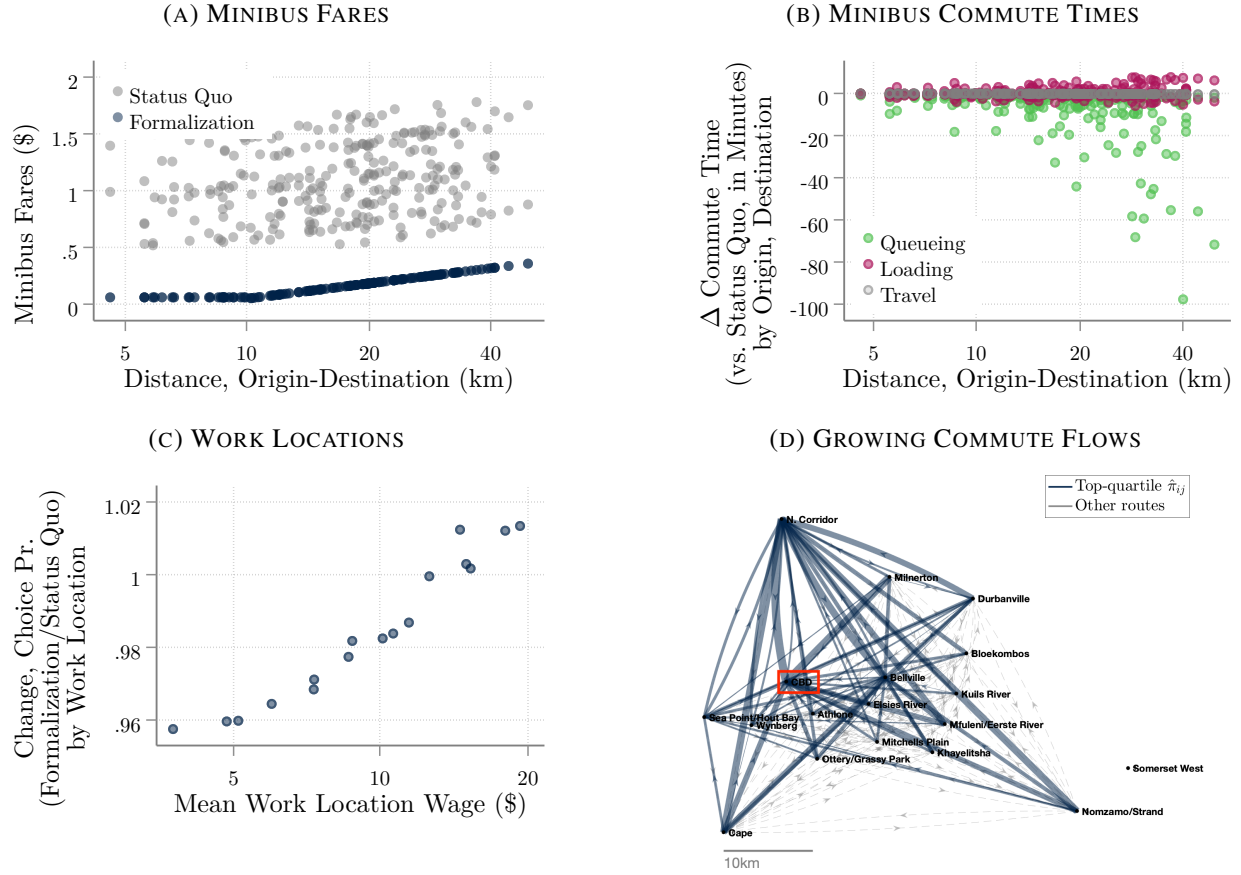
---

taking the total subsidies paid out on a route, multiplying by a skill group's share of minibuses commuters on that route, and summing across routes. Note that, given my focus on feasible policies which target the privatized shared transit sector, I do not include the optimal congestion taxes on car commuters.

<sup>39</sup>Recall that I spell out the back-of-the-envelope calculations and valuations of the implied reductions in greenhouse gas emissions in Online Appendix D.4.8.



**FIGURE 11. MINIBUS FORMALIZATION: LOWER WAIT TIMES AND SPATIAL REALLOCATION**



*Notes:* Figure characterizes formalization counterfactual. Panel (A) displays, on the horizontal axis, the straight-line distance from a minibus route’s origin to destination, and on the vertical axis, the model-predicted status quo minibus fares and the government-set fare scheme in the formalization counterfactual. Panel (B)’s vertical axis instead displays a scatterplot of route-level raw changes in queueing, loading, and travel times,  $Q_{ij}$ ,  $L_{ij}$ , and  $T_{ij}$ , again versus route distance. Panel (C) displays a scatterplot of the proportionate change in work location choice probability, averaged over skill groups, versus the location’s wage, again calculated as the average across skill groups, weighted by aggregate populations. Changes are calculated from the status quo to the formalization counterfactual. Panel (D) maps the 25% of home-work location pairs with the largest proportionate increases in choice probabilities, from the status quo to the formalization counterfactual. Line width corresponds to the percent increase, and CBD is outlined in red.

Effect, yet, as I demonstrate in Online Appendix Table A.1, achieves 94% and 81% of low- and high-skill workers’ gains, respectively, from the full formalization program. The latter “subsidies only” scheme adjusts only the amount of Bus Supply Insurance, which generates less than 20% of the total gains, depending on the skill group. Thus, the Market Mohring Effect drives the gains from formalization, but the fact that the combined gains exceed those from optimal fares alone highlights an important complementarity between these two sources of increasing returns.

### Robustness

In Online Appendix E.3, I explore how four modifications of my model affect the gains from this primary formalization counterfactual. In particular, I simulate the formalization fares and subsidies under (i) nested

logit demand; (ii) lower “non-rush-hour” passenger inflows; (iii) endogenous bus departure timing; and (iv) agglomeration spillovers from local employment to wages. In the first three cases, the gains remain qualitatively similar; agglomeration noticeably increases the benefits of formalization as workers concentrate in ex-ante high-wage locations and thus further push up earnings.

## Enforcement

Next, I evaluate enforcement-related policies aimed at commuters’ road safety and crime-related concerns. Namely, I simulate the enforcement of speed limits for minibuses as well as the government provision of security guards at minibus stations.<sup>40</sup> In each case, I adjust the minibus utility cost  $\kappa_M^g$  by the binary policy effect  $\xi^g$  estimated from the stated preference data; in the case of security, commuters pay their lump-sum share of guard wages.<sup>41</sup> Both of these enforcement-related policies go hand-in-hand with sizable mode shifts towards minibuses. In Figure 12, I focus on the case of station security guards. Higher demand fills buses faster – another manifestation of the Market Mohring Effect, highlighted in the scatterplot of wait and travel times versus route distance in Panel A – and on-road travel times again marginally decrease. The locations that benefit from these additional indirect gains tend to offer lower wages, so in Panel B, commuters reallocate away from the highest-wage work destinations. In Online Appendix Figure A.5, I highlight very similar equilibrium effects of speed limit enforcement. However, the utility and emissions gains of both policies far outweigh the slight decreases in average incomes, so the welfare benefits of these simple interventions, in Table 5, exceed those of any other policy.

## Alternative Policies

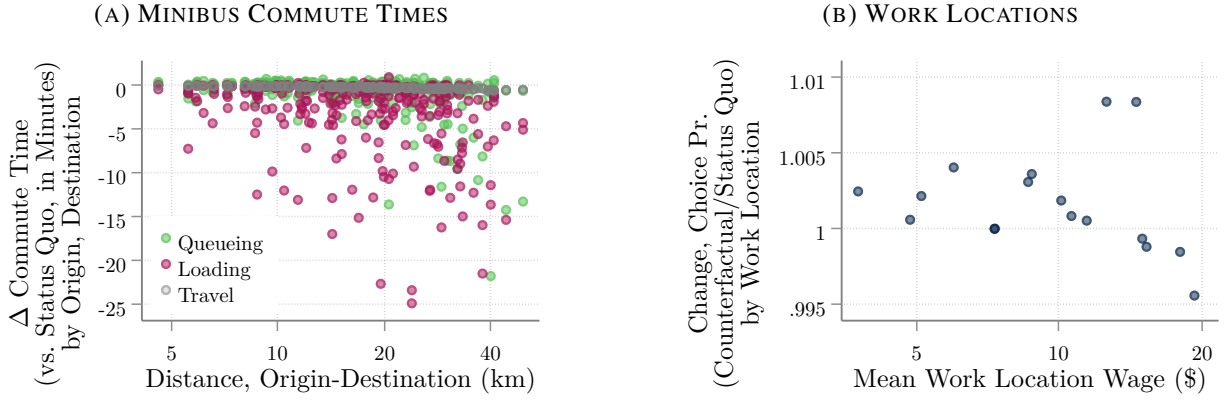
Finally, I consider three alternative policies that mimic other broad classes of privatized shared transit reforms. First, as an alternative strategy to reduce wait times, I simulate imposition of a schedule on routes with below-median passenger arrivals  $\lambda_{ij}$  in the status quo. I require minibus associations on these routes to “schedule” a supply  $b_{ij}$  of buses such that the average passenger faces no queueing time,  $Q_{ij} = 0$ . I further assume that passengers coordinate their arrivals with the schedule, so that they wait only the time it takes the entire busload to board, namely  $L_{ij} = \bar{\eta}/\mu$ .<sup>42</sup> The newly-scheduled routes do indeed experience stark decreases in wait times, but because fares, unlike in my formalization program, do not appreciably decrease, few additional passengers take advantage of these lower waits, as Online Appendix Figures A.6a-A.6b demonstrate. Thus, the welfare gains, in Table 5, turn out marginal – even though I have not even attempted to account for the welfare costs commuters incur when they rearrange their daily routines to match the schedule.

<sup>40</sup>Note that, in the speed limit counterfactual, I do not adjust travel times  $T_{ij}$ , as my calibrated free-flow travel times and estimated road congestion elasticity reflect average traffic speeds rather than minibus-specific speeds more likely to reflect any status-quo speeding. As for the security guard counterfactual, anecdotal 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.

<sup>41</sup>The 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.

<sup>42</sup>Specifically, associations on these routes no longer have a choice of bus supply; instead, they must set bus supply such that the average time in which a busload of passengers arrives equals the average time between bus departures from the station, or  $\bar{\eta}/\lambda_{ij} = L_{ij} + 1/\lambda_{ij}^B = \bar{\eta}/\mu + 1/\lambda_{ij}^B$ , where  $\lambda_{ij}^B$  is still determined by Equation (1).

**FIGURE 12. MINIBUS STATION SECURITY**



*Notes:* Figure characterizes station security counterfactual. Panel (A) displays, on the horizontal axis, the straight-line distance from a minibus route's origin to destination, and on the vertical axis, a scatterplot of route-level raw changes in queueing, loading, and travel times,  $Q_{ij}$ ,  $L_{ij}$ , and  $T_{ij}$ . Panel (B) displays a scatterplot of the proportionate change in work location choice probability, averaged over skill groups, versus the location's wage, again calculated as the average across skill groups, weighted by aggregate populations. Changes are calculated from the status quo to the station security counterfactual.

Second, I simulate the construction of Cape Town's one existing *MyCiti* BRT line to the northern suburbs. To investigate whether this new commute option justifies the construction cost of around \$250 million, I start with a pre-policy economy, for which I predict formal transit wait and travel times  $H_{ij}$  and  $T_{ijF}$  without *MyCiti*.<sup>43</sup> I then reimpose the status-quo formal transit wait and travel times, which naturally include *MyCiti*, but commuters pay annualized construction plus operations costs via equal lump-sum taxes.<sup>44</sup>

The limited spatial reach of the *MyCiti* infrastructure severely limits any direct travel-time gains. Only a handful of home and work locations, displayed in Online Appendix Figure A.6c, benefit from significant inflows of residents and workers. Furthermore, the Market Mohring Effect now comes back to haunt commuters: minibus wait times rise, albeit marginally, as the scale of minibus demand falls. Wait times increase the most not on routes which directly compete with *MyCiti*, but rather, as depicted in Online Appendix Figure A.6d, on routes far from the *MyCiti* corridor, which already had the longest waits. In turn, the fixed costs of *MyCiti* significantly exceed any benefits, so net welfare, in Table 5, decreases. Indeed, in sprawling cities like Cape Town, widely dispersed home and work locations may not justify costly formal transit infrastructure.

<sup>43</sup>Specifically, lacking data on pre-*MyCiti* formal transit wait and travel times, I build a formal transit network analogous to the road network, where only neighboring locations are linked. Total time to traverse a link equals the respective formal transit wait plus travel time using the same data described in Online Appendix D.4.5. I calculate the shortest path through this network and the corresponding total wait and total travel time. I then remove the Northern Suburbs-CBD formal transit network link which corresponds to *MyCiti* and recalculate total wait and travel time between each set of origins and destinations. For each origin and destination, I calculate the ratio between the latter and the former total commute times, i.e. without and with *MyCiti*, and then multiply this ratio times the baseline  $H_{ij}$  and  $T_{ijF}$  to obtain the pre-*MyCiti* wait and travel times.

<sup>44</sup>*MyCiti* costs derive from the City of Cape Town (2015). Construction costs include R4,157,851,000 in Phase 1 infrastructure, vehicle, and compensation costs (2006-2013) in Table 9-1 plus R1,062,320 in planning and transition costs (2006-2013) in Table 9-3, which I annualize with an interest rate of 7% and a lifespan of 60 years and divide by the number of (half) working days per year. Operating costs per-commute are R390,447,000 in "deficit before funding," i.e. net of fare revenue, in 2015/16, from Table 9-4, divided by the number of (half) working days per year.

**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		0.05	0.04	-0.01	-0.03	0.38	-5.29	0.30	0.14	0.38	0.17
Formalization		0.05	0.04	-0.01	-0.02	0.45	-3.67	0.30	0.13	0.35	0.16
Speed Limit		0.06	0.03	-0.01	-0.01	-0.06	-3.56	4.45	0.76	4.50	0.78
Station Security		0.05	0.10	-0.01	-0.05	-0.07	-5.17	3.41	2.87	3.49	2.90
Minibus Schedule		0.001	0.004	-0.0002	-0.002	-0.06	-0.13	0.06	0.1	0.06	0.1
MyCiti BRT		-0.001	-0.002	-0.001	-0.004	0.04	-0.36	-0.91	-0.33	-0.9	-0.33
Larger Minibuses		-0.0003	-0.002	0.0001	0.0009	0.02	0.08	-0.02	-0.04	-0.02	-0.04

*Notes:* Table summarizes the social planner optimum and each counterfactual policy: a minibus formalization program which approximates the social optimum via optimal fares and association subsidies, speed limit enforcement for minibuses, adding security guards to all minibus stations, imposing a schedule on the bottom half of minibus routes by status-quo demand, construction of Cape Town’s MyCiti bus rapid transit, and a larger minibus recapitalization program which increases minibus capacity  $\bar{\eta}$  to 23. 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.4.8. 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. Note that all changes are taken relative to the status quo, except in the case of MyCiti, where I compute changes from the pre-MyCiti equilibrium to a status quo that accounts for MyCiti’s construction and operations costs.

Third, I simulate a hypothetical second-round Minibus Taxi Recapitalization program, which additionally stipulates that larger, 23-seat buses replace the current 15-seat version. I abstract away from the transfers necessary to induce this switch and instead, for simplicity, simulate an across-the-board increase in bus size to  $\bar{\eta}' = 23$ .<sup>45</sup> Larger buses, to the extent that they continue to depart full, will mechanically increase passengers’ total wait times for given rates of passenger arrivals. Moreover, the model highlights an important feedback effect: longer wait times reduce equilibrium minibus demand and thus further push up loading times, particularly on the longest routes, as in Online Appendix Figures A.6e-A.6f. Importantly, on-road travel times, also in Online Appendix Figure A.6f, scarcely change; evidently, minibuses do not constitute a sufficient fraction of total traffic to substantially diminish road congestion. Thus, far from de-congesting Cape Town’s busy streets, a policy which increases bus size would, in Table 5, marginally decrease welfare.

## VIII. CONCLUSION

In this paper, I build the first model of the privatized shared transit sector which dominates many developing-country cities. Unlike existing models, which take commute costs as given, mine predicts how wait times, travel times, and fares respond to the choices of passengers and transit providers in equilibrium. In particular, my queueing model generates a market-driven version of the Mohring (1972) increasing returns in wait

<sup>45</sup>Since I lack reliable operations cost data on hypothetical smaller minibuses, I adjust  $\chi_{ij}$  by an equivalent factor.

times inherent to all forms of shared transit. The strict capacity constraint of minibuses, however, introduces a new role for surplus bus supply to prevent the growth of passenger queues. Market power inhibits the realization of the former Market Mohring Effect but facilitates the internalization of the latter Bus Supply Insurance Effect. I collect new data on passenger queues and bus arrivals in Cape Town to directly estimate key queueing and bus arrival efficiency parameters. Furthermore, I introduce the stated preference approach to the urban literature to identify the commuter demand system. Finally, I quantify the scope for low-cost interventions that ameliorate the wait time and safety frictions which loom large in Cape Town’s current privatized shared transit network.

Three policies’ welfare and equity effects stand out. A minibus formalization program sets fares optimally to increase the scale of minibus commuting and leverage the Market Mohring Effect, but simultaneously subsidizes associations to maintain sufficient Bus Supply Insurance. The associated decreases in queueing and loading times allow commuters to work in higher-wage locations and generate welfare gains that rival those of previously-studied urban transit infrastructure investments. Minibus speed limit enforcement or station security guards prove similarly beneficial, in part because they indirectly decrease wait times in a Market-Mohring fashion. All three policies decrease road congestion as well as carbon emissions and disproportionately benefit low-skill commuters. Even absent the fiscal capacity for more substantial infrastructure investments, policymakers thus enjoy ample scope to improve upon the privatized provision of transit.

## APPENDIX

### A. THEORY

#### A.1 Queue Simulation

In this section, I discuss how I simulate my queueing model out of steady state. In particular, consider how the queue for a given route  $ij$  evolves in continuous time over some instants indexed by  $t$ . For readability, I suppress the  $ij$ -subscript throughout this section. Denote the (joint) probability that, at instant  $t$ ,  $n$  passengers wait in the queue and a bus is present as  $p_n^b(t)$  and the probability that  $n$  passengers wait in the queue and *no* bus is present as  $q_n(t)$ .

As in Brancaccio et al. (2024), each of these probabilities evolves according to Kolmogorov Forward Equations, as follows. For the pdf of queue lengths with a bus present:

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

For interpretation, consider the case of no queue but a bus present and loading, where  $n = 0$ . Over time, this probability decreases by an “outflow” in the first term equal to the probability of that state times the rate  $\lambda$  at which another passenger arrives, taking the queue into state ( $n = 1$ , bus present), plus the (time-varying) rate

$\mu^B(t)$  at which the bus departs, taking the queue into state  $(n = 0, \text{no bus})$ . The probability  $p_0^b(t)$  increases by the probability of the state  $(n = 0, \text{no bus})$  times the rate  $\lambda^B$  at which buses arrive plus the probability of a queue of 1 with a bus present times the rate  $\mu$  at which that passenger boards the bus. The evolution of  $p_n^b(t)$  for  $n > 0$  must additionally account for an outflow of passengers boarding that bus at rate  $\mu$  as well as an increase over time in this probability by the probability of a one-passenger-shorter queue times the rate at which an additional passenger arrives.

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

$$\frac{d}{dt}q_n(t) = \begin{cases} -(\lambda + \lambda^B)q_0(t) + \mu^B(t)p_0^b(t), & \text{if } n = 0 \\ -(\lambda + \lambda^B)q_n(t) + \mu^B(t)p_n^b(t) + \lambda q_{n-1}(t), & \text{if } n > 0 \end{cases} \quad (\text{A.2})$$

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

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

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

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

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

and

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

By substituting the given  $\lambda^B$  and the (expected)  $L$  obtained through this simulation into Equation (1), I obtain the necessary bus supply  $b$ .<sup>46</sup>

---

<sup>46</sup>For computational reasons, I set a maximum queue length  $N$ , sufficiently high such that  $p_N^b(t)$  and  $q_N(t)$  are small, and also solve only for every 15th  $p_n^b(t)$  and  $q_n(t)$  for  $n$  above some threshold  $\tilde{N}$ . I adjust the Kolmogorov Forward Equations (A.1)-(A.2)

## A.2 Efficiency Proofs

I first derive a conveniently rewritten form of the welfare function  $\Omega$ .

**Lemma A.1.** *Welfare satisfies*

$$\begin{aligned} \Omega = \sum_{i,j,g} N^g \pi_{ijM}^g & \left[ \theta_i^g - \kappa_M^g - r\omega_j^g (Q_{ij} + L_{ij} + T_{ij}) + \omega_j^g - v \log \pi_{ijM}^g - \frac{\chi_{ij}}{\bar{\eta}} \right] \\ & + \sum_{i,j,g} N^g \pi_{ijF}^g \left[ \theta_i^g - \kappa_F^g - r\omega_j^g (H_{ij} + T_{ijF}) - \tau_{ijF} + \omega_j^g - v \log \pi_{ijF}^g \right] \\ & + \sum_{i,j,g} N^g \pi_{ijA}^g \left[ \theta_i^g - \kappa_A^g - r\omega_j^g T_{ij} - \tau_A + \omega_j^g - v \log \pi_{ijA}^g \right] - \bar{\omega} \sum_{i,j} b_{ij}. \quad (\text{A.3}) \end{aligned}$$

*Proof.* In (11), consider first the ex-ante expected utility of group- $g$  commuters,  $\bar{\Omega}^g$ , given their optimal choices of home, work, and mode, subject to idiosyncratic Gumbel-distributed preference shocks.<sup>47</sup> 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

$$\begin{aligned} \bar{\Omega}^g & \equiv E \left[ \max_{i',j',m'} (\bar{U}_{i'j'm'}^g + v \varepsilon_{i'j'm'}) \right] \\ & = \sum_{i,j,m} \pi_{ijm}^g [\bar{U}_{ijm}^g + v E(\varepsilon_{ijm} | ijm \in \arg\max_{i',j',m'} (\bar{U}_{i'j'm'}^g + v \varepsilon_{i'j'm'}))] \\ & = \sum_{i,j,m} \pi_{ijm}^g [\theta_i^g + U_{ijm}^g + \omega_j^g - v \log \pi_{ijm}^g]. \quad (\text{A.4}) \end{aligned}$$

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.<sup>48</sup> Next, I substitute (10) into association profits (5) and sum across routes to obtain  $\Pi \equiv \sum_{i,j} \Pi_{ij}$ . Finally, substituting each element into (11), using the mode-specific definitions of commute utility  $U_{ijm}^g$ , and rearranging, I obtain the expression in (A.3).  $\square$

### Proof of Proposition 1

*Proof.* First, I derive the conditions characterizing the social planner optimum. From the definition of

accordingly.

<sup>47</sup>Note that the social planner will choose choice probabilities directly and then implement these choice probabilities with appropriately-set transfers, as in Appendix A.3.

<sup>48</sup>To see this, note that

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

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  $\bar{\Omega}^g$ , and the final uses the choice probability equation (9).



optimality in the main text, the planner solves

$$\max_{\{\pi\}_{i,j,m,g}\{b\}_{i,j}} \Omega \text{ s.t. } Q_{ij} = Q(\lambda_{ij}, b_{ij}), L_{ij} = L(\lambda_{ij}, b_{ij}), T_{ij} = \sum_{kk' \in s(i,j)} \bar{t}_{kk'} x_{kk'}^\gamma, \text{ and } \sum_{i,j,m} \pi_{ijm}^g = 1. \quad (\text{A.5})$$

I substitute the  $Q(\cdot)$  and  $L(\cdot)$  which result from my numerical queue simulations as well as the expression for link-level vehicle volume,  $x_{kk'} = \sum_{i',j':kk' \in s(i',j')} \sum_{g'} N^{g'} \left( \pi_{i'j'M}^{g'}/\bar{\eta} + \pi_{i'j'A}^{g'} \right)$ , into the expression in Lemma A.1 to obtain welfare as  $\Omega(b, \pi)$ . I can rewrite the planner's problem as  $\max_{b, \pi} \Omega(b, \pi)$  s.t.  $\sum_{i,j,m} \pi_{ijm}^g = 1$ .<sup>49</sup> The planner's first-order conditions for bus supply  $b_{ij}$  on each route reads

$$\frac{\partial \Omega}{\partial b_{ij}} = - \sum_g N^g \pi_{ijM}^{g*} r \omega_j^g \left[ \frac{\partial Q_{ij}}{\partial b_{ij}} + \frac{\partial L_{ij}}{\partial b_{ij}} \right] - \bar{\omega} = 0. \quad (\text{A.6})$$

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

$$\begin{aligned} \log \left( \pi_{ijm}^{g*} / \pi_{klF}^{g*} \right) = & \exp \left( \theta_i^g - \kappa_m^g + \omega_j^g + 1 \{m = M\} \left\{ -r \omega_j^g (Q_{ij} + L_{ij} + T_{ij}) - \frac{\chi_{ij}}{\bar{\eta}} \right. \right. \\ & - \sum_{g'} N^{g'} \pi_{ijM}^{g'*} r \omega_j^{g'} \left( \frac{\partial Q_{ij}}{\partial \lambda_{ij}} + \frac{\partial L_{ij}}{\partial \lambda_{ij}} \right) - \sum_{i',j',g'} N^{g'} \left( \pi_{i'j'M}^{g'*} + \pi_{i'j'A}^{g'*} \right) r \omega_{j'}^{g'} \frac{\partial T_{i'j'}}{\partial \pi_{ijM}^g} \frac{1}{N^g} \left. \right\} \\ & + 1 \{m = A\} \left[ -r \omega_j^g T_{ij} - \tau_A - \sum_{i',j',g'} N^{g'} \left( \pi_{i'j'M}^{g'*} + \pi_{i'j'A}^{g'*} \right) r \omega_{j'}^{g'} \frac{\partial T_{i'j'}}{\partial \pi_{ijA}^g} \frac{1}{N^g} \right] \\ & + 1 \{m = F\} \left[ -r \omega_j^g (H_{ij} + T_{ijF}) - \tau_{ijF} \right] - \theta_k^g + \kappa_F^g + r \omega_l^g (H_{kl} + T_{klF}) + \tau_{klF} - \omega_l^g \Big)^{1/\nu}. \quad (\text{A.7}) \end{aligned}$$

Note that I calculate  $\frac{\partial Q_{ij}}{\partial b_{ij}}, \frac{\partial L_{ij}}{\partial b_{ij}}, \frac{\partial Q_{ij}}{\partial \lambda_{ij}}$ , and  $\frac{\partial L_{ij}}{\partial \lambda_{ij}}$  numerically and  $\frac{\partial T_{i'j'}}{\partial \pi_{ijM}^g}$  as well as  $\frac{\partial T_{i'j'}}{\partial \pi_{ijA}^g}$  using the chain rule and the expressions introduced in the main text. Equations (A.6), (A.7), and the adding-up constraints in (A.5) 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  $\pi_{ijm}^{g*}$ .

Second, I derive conditions under which the associations and commuters internalize the potential sources of inefficiency, as defined in the main text. Starting with bus supply, equilibrium supply  $b_{ij}$  coincides with the optimal entry  $b_{ij}^*$  pinned down by the planner's bus entry condition (A.6) if and only if

$$\left( \tau_{ij} - \frac{\chi_{ij}}{\bar{\eta}} \right) \sum_g N^g \frac{\partial \pi_{ijM}^g}{\partial b_{ij}} = - \sum_g N^g \pi_{ijM}^{g*} r \omega_j^g \left[ \frac{\partial Q_{ij}}{\partial b_{ij}} + \frac{\partial L_{ij}}{\partial b_{ij}} \right]. \quad (\text{A.8})$$

Rearranging yields (13) in the main text.

<sup>49</sup>This welfare function includes the expectation of the idiosyncratic shocks, under the assumption that the planner induces the optimal share of commuters of each skill group to choose each home, work, and mode tuple through through appropriately-set transfers, as in Appendix A.3.

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

$$\begin{aligned} \log \left( \pi_{ijm}^{g*|T} / \pi_{klF}^{g*|T} \right) = & \exp \left( \theta_i^g - \kappa_m^g + \omega_j^g + 1 \{m = M\} \left\{ -r\omega_j^g (Q_{ij} + L_{ij} + T_{ij}) - \frac{\chi_{ij}}{\bar{\eta}} \right. \right. \\ & \left. \left. - \sum_{g'} N^{g'} \pi_{ijM}^{g'*|T} r\omega_j^{g'} \left( \frac{\partial Q_{ij}}{\partial \lambda_{ij}} + \frac{\partial L_{ij}}{\partial \lambda_{ij}} \right) \right\} + 1 \{m = A\} \left[ -r\omega_j^g T_{ij} - \tau_A \right] \right. \\ & \left. + 1 \{m = F\} \left[ -r\omega_j^g (H_{ij} + T_{ijF}) - \tau_{ijF} \right] - \theta_k^g + \kappa_F^g + r\omega_l^g (H_{kl} + T_{klF}) + \tau_{klF} - \omega_l^g \right)^{1/\nu}. \quad (\text{A.9}) \end{aligned}$$

I substitute commute values  $U_{ijm}^g$  into (9); it is then immediate that equilibrium (relative) choice probabilities,  $\log \left( \pi_{ijm}^g / \pi_{klF}^g \right)$ , equal those chosen by the planner, holding travel times fixed, in (A.9) if and only if bargained minibus fares compensate, for each route and skill group, per-passenger operations costs plus the marginal value of the Market Mohring Effect, as in (12).

In a similar vein, I take the planner's first-order conditions with respect to commuter choice probabilities  $\pi_{ijm}^g$ , holding minibus wait times  $\{Q, L\}$  fixed, to correspond to the definition of the Road Congestion Externality. I again derive an optimal relative choice probability for each origin-by-destination-by-mode-by-skill, holding  $\{Q, L\}$  fixed:

$$\begin{aligned} \log \left( \pi_{ijm}^{g*|Q,L} / \pi_{klF}^{g*|Q,L} \right) = & \exp \left( \theta_i^g - \kappa_m^g + \omega_j^g + 1 \{m = M\} \left\{ -r\omega_j^g (Q_{ij} + L_{ij} + T_{ij}) - \frac{\chi_{ij}}{\bar{\eta}} \right. \right. \\ & \left. \left. - \sum_{i',j',g'} N^{g'} \left( \pi_{i'j'M}^{g'*|Q,L} + \pi_{i'j'A}^{g'*|Q,L} \right) r\omega_{j'}^{g'} \frac{\partial T_{i'j'}}{\partial \pi_{ijM}^g} \frac{1}{N^g} \right\} \right. \\ & \left. + 1 \{m = A\} \left[ -r\omega_j^g T_{ij} - \tau_A - \sum_{i',j',g'} N^{g'} \left( \pi_{i'j'M}^{g'*|Q,L} + \pi_{i'j'A}^{g'*|Q,L} \right) r\omega_{j'}^{g'} \frac{\partial T_{i'j'}}{\partial \pi_{ijA}^g} \frac{1}{N^g} \right] \right. \\ & \left. + 1 \{m = F\} \left[ -r\omega_j^g (H_{ij} + T_{ijF}) - \tau_{ijF} \right] - \theta_k^g + \kappa_F^g + r\omega_l^g (H_{kl} + T_{klF}) + \tau_{klF} - \omega_l^g \right)^{1/\nu}. \quad (\text{A.10}) \end{aligned}$$

Comparison with the equilibrium relative choice probabilities based on (9) makes it immediate that car choice probabilities never equal these constant-wait-time optimal probabilities, i.e.  $\pi_{ijA}^g \neq \pi_{ijA}^{g*|Q,L}$ . Note that minibus commuters could in theory internalize their road congestion spillovers in another knife-edge case,  $\tau_{ij} = \frac{\chi_{ij}}{\bar{\eta}} + \sum_{i',j',g'} N^{g'} \left( \pi_{i'j'M}^{g'*|Q,L} + \pi_{i'j'A}^{g'*|Q,L} \right) r\omega_{j'}^{g'} \frac{\partial T_{i'j'}}{\partial \pi_{ijM}^g} \frac{1}{N^g}$ . However, commuters will never universally internalize the Road Congestion Externality.  $\square$

### A.3 Decentralization of Social Planner Optimum

In this section, I derive the minibus fares  $\tau_{ij}$ , per-passenger subsidies to associations  $z_{ij}$ , as well as congestion taxes  $t_{ij}$  on car commuters that induce the socially-optimal bus supply  $b_{ij}^*$  and commuter choices  $\pi_{ijm}^{g*}$ . First, consider minibus commuter choice probabilities. By comparing the social planner-optimal choice probabilities (A.7) with the equilibrium (relative) choice probabilities implied by (9), we can see that the optimal minibus choice probabilities can be implemented by the fares

$$\tau_{ij} = \frac{\chi_{ij}}{\eta} + \sum_{g'} N^{g'} \pi_{ijM}^{g'*} r \omega_j^{g'} \left( \frac{\partial Q_{ij}}{\partial \lambda_{ij}} + \frac{\partial L_{ij}}{\partial \lambda_{ij}} \right) + \sum_{i',j',g'} N^{g'} \left( \pi_{i'j'M}^{g'*} + \pi_{i'j'A}^{g'*} \right) r \omega_{j'}^{g'} \frac{\partial T_{i'j'}}{\partial \lambda_{ij}}. \quad (\text{A.11})$$

where I use the fact that  $\frac{\partial T_{i'j'}}{\partial \pi_{ijM}^{g'}} \frac{1}{N^g} \equiv \frac{\partial T_{i'j'}}{\partial \lambda_{ij}}$ .

Next, consider bus supply,  $b_{ij}$ . From Proposition 1, associations internalize the Bus Supply Insurance Effect when they receive a fare—gross of any per-passenger subsidies, which, of course, affect association profits in the same manner as fares—as given by Equation (13). Thus, the per-passenger subsidy that induces associations to choose efficient bus supply equals the gross fares given by (13) minus the net fare in Equation A.11, or

$$z_{ij} = - \left( \frac{\partial \lambda_{ij}}{\partial b_{ij}} \right)^{-1} \sum_{g'} N^{g'} \pi_{ijM}^{g'*} r \omega_j^{g'} \left( \frac{\partial Q_{ij}}{\partial b_{ij}} + \frac{\partial L_{ij}}{\partial b_{ij}} \right) - \sum_{g'} N^{g'} \pi_{ijM}^{g'*} r \omega_j^{g'} \left( \frac{\partial Q_{ij}}{\partial \lambda_{ij}} + \frac{\partial L_{ij}}{\partial \lambda_{ij}} \right) - \sum_{i',j',g'} N^{g'} \left( \pi_{i'j'M}^{g'*} + \pi_{i'j'A}^{g'*} \right) r \omega_{j'}^{g'} \frac{\partial T_{i'j'}}{\partial \lambda_{ij}}. \quad (\text{A.12})$$

Finally, I derive the congestion taxes  $t_{ij}$  on car commuters which induce the optimal car choice probabilities  $\pi_{ijA}^{g*}$ . Taxes  $t_{ij}$  are paid along with the fare upon choice of a commute. To highlight the fact that taxes do not differ by skill group, I define the inflow of car commuters as  $c_{ij} \equiv \sum_{g'} N^{g'} \pi_{ijA}^{g'}$  such that, as above in the minibus case,  $\frac{\partial T_{i'j'}}{\partial \pi_{ijA}^{g'}} \frac{1}{N^g} \equiv \frac{\partial T_{i'j'}}{\partial c_{ij}}$ . By again comparing the social planner-optimal choice probabilities (A.7) with the equilibrium (relative) choice probabilities implied by (9), we can see that the optimal car choice probabilities can be implemented by taxes given by

$$t_{ij} = \sum_{i',j',g'} N^{g'} \left( \pi_{i'j'M}^{g'*} + \pi_{i'j'A}^{g'*} \right) r \omega_{j'}^{g'} \frac{\partial T_{i'j'}}{\partial c_{ij}}. \quad (\text{A.13})$$

## REFERENCES

- Abenzoza, Roberto F., Oded Cats, and Yusak O. Susilo. 2019. “How does travel satisfaction sum up? An exploratory analysis in decomposing the door-to-door experience for multimodal trips.” *Transportation* 46:1615–1642.
- Ahlfeldt, Gabriel M., Stephen J. Redding, Daniel M. Sturm, and Nikolaus Wolf. 2015. “The Economics of Density: Evidence from the Berlin Wall.” *Econometrica* 83 (6): 2127–2189.

- Almagro, Milena, Felipe Barbieri, Juan Camilo Castillo, Nathaniel Hickok, and Tobias Salz. 2024. “Optimal Urban Transportation Policy: Evidence from Chicago.”
- Ansari Esfeh, Mohammad, S. C. Wirasinghe, Saeid Saidi, and Lina Kattan. 2021. “Waiting time and headway modelling for urban transit systems – a critical review and proposed approach.” *Transport Reviews* 41 (2): 141–163.
- Antrobus, Lauren, and Andrew Kerr. 2019. “The labour market for minibus taxi drivers in South Africa.” SALDRU Working Paper No. 250.
- Arroyo-Arroyo, Fatima, Antoine Chevre, Pablo Salazar Ferro, Julien Allaire, and Marion Hoyez. 2021. *Studies of Informal Passenger Transport Reforms in Sub-Saharan Africa: Senegal: Dakar*. Technical report. SSATP, July. [https://www.ssatp.org/sites/default/files/publication/Dakar\\_vf.pdf](https://www.ssatp.org/sites/default/files/publication/Dakar_vf.pdf).
- Arroyo-Arroyo, Fatima, Antoine Chevre, Herrie Schalekamp, Nico McLachlan, Marion Hoyez, and Pablo Salazar Ferro. 2021. *Studies of Informal Passenger Transport Reforms in Sub-Saharan Africa: South Africa: Cape Town*. Technical report. SSATP, July. [https://www.ssatp.org/sites/default/files/publication/Cape\\_Town\\_vf.pdf](https://www.ssatp.org/sites/default/files/publication/Cape_Town_vf.pdf).
- 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.”
- Barrett, Ian, Brendan Finn, and Xavier Godard. 2016. “West African case studies of integrated urban transport reform.” In *Paratransit in African Cities: Operations, Regulation and Reform*, edited by Roger Behrens, Dorothy McCormick, and David Mfinanga, 244–271. Routledge.
- 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*.
- Behrens, Roger, and Pablo Salazar Farro. 2016. “Barriers to comprehensive paratransit replacement.” In *Paratransit in African Cities: Operations, Regulation and Reform*, edited by Roger Behrens, Dorothy McCormick, and David Mfinanga, 199–220. Routledge.
- Behrens, Roger, Pablo Salazar Farro, and Aaron Golub. 2016. “International case studies of hybrid public transport system regulation and complementarity.” In *Paratransit in African Cities: Operations, Regulation and Reform*, edited by Roger Behrens, Dorothy McCormick, and David Mfinanga, 221–243. Routledge.
- Ben-Akiva, Moshe, Daniel McFadden, and Kenneth Train. 1919. “Foundations of Stated Preference Elicitation: Consumer Behavior and Choice-based Conjoint Analysis.” *Foundations and Trends in Econometrics* 10 (1–2): 1–144.
- Björkegren, Daniel, Alice Duhaut, Geetika Nagpal, and Nick Tsivanidis. 2025. “Public and Private Transit: Evidence from Lagos.”
- 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 Kalouptsi, and Theodore Papageorgiou. 2024. “Investment in Infrastructure and Trade: The Case of Ports.”
- Bryan, Gharad, Edward Glaeser, and Nick Tsivanidis. 2020. “Cities in the Developing World.” *Annual Review of Economics* 12:21.1–21.25.

- Bryan, Gharad, and Melanie Morten. 2019. “The Aggregate Productivity Effects of Internal Migration: Evidence from Indonesia.” *Journal of Political Economy* 127 (5): 2229–2268.
- Buchholz, Nicholas, Laura Doval, Jakub Kastl, Filip Matejka, and Tobias Salz. 2024. “Personalized Pricing and the Value of Time: Evidence from Auctioned Cab Rides.” *Revise & Resubmit, Econometrica*.
- Cervero, Robert, and Aaron Golub. 2007. “Informal transport: A global perspective.” *Transport policy* 14 (6): 445–457.
- 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>.
- . 2015. *MyCiti Business Plan 2015 Update*. Technical report. Cape Town City Council, March. <https://www.myciti.org.za/docs/categories/1605/MyCiTi%5C%20Business%5C%20Plan%5C%202015%5C%20Update.pdf>.
- Coetzee, Justin, Christoff Krogscheepers, and John Spotten. 2018. “Mapping minibus-taxi operations at a metropolitan scale - methodologies for unprecedented data collection using a smartphone application and data management techniques.”
- Collard-Wexler, Allan, Gautam Gowrisankaran, and Robin S. Lee. 2019. “Nash-in-Nash” Bargaining: A Microfoundation for Applied Work.” *Journal of Political Economy* 127 (1): 163–195.
- Combes, Pierre-Philippe, and Laurent Gobillon. 2015. “The Empirics of Agglomeration Economies.” In *Handbook of Regional and Urban Economics*, vol. 5A, 247–348. Elsevier.
- De Vos, Jonas, Alireza Ermagun, and F. Atiyya Shaw. 2023. “Wait time, travel time and waiting during travel: existing research and future directions.” *Transport Reviews* 43 (5): 805–810.
- Filho, Romulo Dante Orrico, Renato Guimaraes Ribeiro, and Mame Khadidiatou Thiam. 2015. “A comparative study of the organization of alternative transport in the cities of Rio de Janeiro and Dakar.” *Case Studies on Transport Policy* 3:278–284.
- Gechter, Michael, and Nick Tsivanidis. 2023. “Spatial Spillovers from High-Rise Developments: Evidence from the Mumbai Mills.” *Revise & Resubmit accepted, Econometrica*.
- Goldszmidt, Ariel, John A. List, Robert D. Metcalfe, Ian Muir, V. Kerry Smith, and Jenny Wang. 2020. “The Value of Time in the United States: Estimates from Nationwide Natural Field Experiments.”
- Harari, Mariaflavia, and Maisy Wong. 2024. “Slum Upgrading and Long-run Urban Development: Evidence from Indonesia.” *Accepted, Review of Economic Studies*.
- Hosios, Arthur J. 1990. “On The Efficiency of Matching and Related Models of Search and Unemployment.” *The Review of Economic Studies* 57 (2): 279–298.
- Hsiao, Allan. 2023. “Sea Level Rise and Urban Adaptation in Jakarta.”
- . 2024. “Educational Investment in Spatial Equilibrium: Evidence from Indonesia.”
- Hurtubia, Ricardo, and Janus Leonhardt. 2021. *The Experience of Reforming Bus Concessions in Santiago de Chile*. Technical report. International Transport Forum. <https://www.itf-oecd.org/sites/default/files/docs/reforming-bus-concessions-santiago-de-chile.pdf>.
- Jennings, Gail, Eric Bruun, Risper Orero, Dorothy McCormick, and Paul Browning. 2016. “Strategy options for paratransit: business development and service improvement.” In *Paratransit in African Cities: Operations, Regulation and Reform*, edited by Roger Behrens, Dorothy McCormick, and David Mfinanga, 272–305. Routledge.
- Jobanputra, Rahul. 2018. *Comprehensive Integrated Transport Plan 2018 – 2023*. Technical report. City of Cape Town Transport and Urban Development Authority, January. <https://tdacontenthubstore.blob.core.windows.net/resources/fd3ddc0d-b459-4d26-bb01-7f689d7a36eb.pdf>.

- Johnston, Robert J., Kevin J. Boyle, Wiktor (Vic) Adamowicz, Jeff Bennett, Roy Brouwer, Trudy Ann Cameron, W. Michael Hanemann, et al. 2017. "Contemporary Guidance for Stated Preference Studies." *Journal of the Association of Environmental and Resource Economists* 4 (2): 319–405.
- Kerr, Andrew. 2018. *Background note: Minibus Taxis, Public Transport, and the Poor*. Technical report. World Bank. <https://openknowledge.worldbank.org/handle/10986/30018>.
- Kerzhner, Tamara. 2022. "Is informal transport flexible?" *The Journal of Transport and Land Use* 15 (1): 671–689.
- . 2023. "How are informal transport networks formed? Bridging planning and political economy of labour." *Cities* 137:104348.
- Khanna, Gaurav, Carlos Medina, Anant Nyshadham, Daniel Ramos-Menchelli, Jorge Tamayo, and Audrey Tiew. 2024. "Spatial Mobility, Economic Opportunity, and Crime." *Revise & Resubmit, American Economic Review*.
- Kreindler, Gabriel E. 2022. "Peak-Hour Road Congestion Pricing: Experimental Evidence and Equilibrium Implications." *Conditionally accepted, Econometrica*.
- Kreindler, Gabriel E., Arya Gaduh, Tilman Graff, Rema Hanna, and Benjamin A. Olken. 2023. "Optimal Public Transportation Networks: Evidence from the World's Largest Bus Rapid Transit System in Jakarta."
- Mangham, Lindsay J., Kara Hanson, and Barbara McPake. 2009. "How to do (or not to do)...Designing a discrete choice experiment for application in a low-income country." *Health Policy and Planning* 24:151–158.
- Mbonu, Oluchi, and F. Christopher Eaglin. 2024. "Market Segmentation and Coordination Costs: Evidence from Johannesburg's Minibus Networks."
- Mfinanga, David, and Erick Mafinda. 2016. "Public transport and daladala service improvement prospects in Dar es Salaam." In *Paratransit in African Cities: Operations, Regulation and Reform*, edited by Roger Behrens, Dorothy McCormick, and David Mfinanga, 155–173. Routledge.
- Miyauchi, Yuhei, Kentaro Nakajima, and Stephen J. Redding. 2022. "The Economics of Spatial Mobility: Theory and Evidence Using Smartphone Data."
- Mohring, Herbert. 1972. "Optimization and Scale Economies in Urban Bus Transportation." *American Economic Review* 62 (4): 591–604.
- Nikolaidou, Anastasia, Aristomenis Kopsacheilis, Georgios Georgiadis, Theodoros Noutsias, Ioannis Politis, and Ioannis Fyrogenis. 2023. "Factors affecting public transport performance due to the COVID-19 outbreak: A worldwide analysis." *Cities* 134:104206.
- Olea, José Luis Montiel, and Carolin Pflueger. 2013. "A Robust Test for Weak Instruments." *Journal of Business Economic Statistics* 31 (3): 358–369.
- Olvera, Lourdes Diaz, Didier Plat, and Pascal Pochet. 2024. "Changes in daily mobility and new public transport supply in Dakar (2000 – 2015)." *Case Studies on Transport Policy* 16:101214.
- Parry, Ian W. H., and Kenneth A. Small. 2009. "Should Urban Transit Subsidies Be Reduced?" *American Economic Review* 99 (3): 700–724.
- Ribbonaar, D., R. Collins, G. Martin, B. Macmahon, G.M. Johnson, L. Rautenbach, P. Grey, M. Moody, and Z. Ahmed. 2023. "Minibus Taxi Improvement Initiatives in the Western Cape."
- Rodriguez, Camila, Tatiana Peralta-Quirós, Luis A. Guzman, and Sebastian A. Cárdenas Reyes. 2017. "Accessibility, Affordability, and Addressing Informal Services in Bus Reform: Lessons from Bogotá, Colombia." *Transportation Research Record* 2634:35–42.
- Romero de Tejada, Joaquín, Anna Mazzolini, Constâncio Machanguana, António Matos and Gécica Macamo, Clemence Cavoli, and Daniel Oviedo. 2023. *Maputo City Profile. Mobility, Accessibility and Land Use in the Maputo Metropolitan Area*. Technical report. Transitions to Sustainable Urban Mobility (T-SUM) Project. [https://www.t-sum.org/\\_files/ugd/b4aba8\\_c1853e0511ee45e0bc5ab0884ff7d24b.pdf](https://www.t-sum.org/_files/ugd/b4aba8_c1853e0511ee45e0bc5ab0884ff7d24b.pdf).



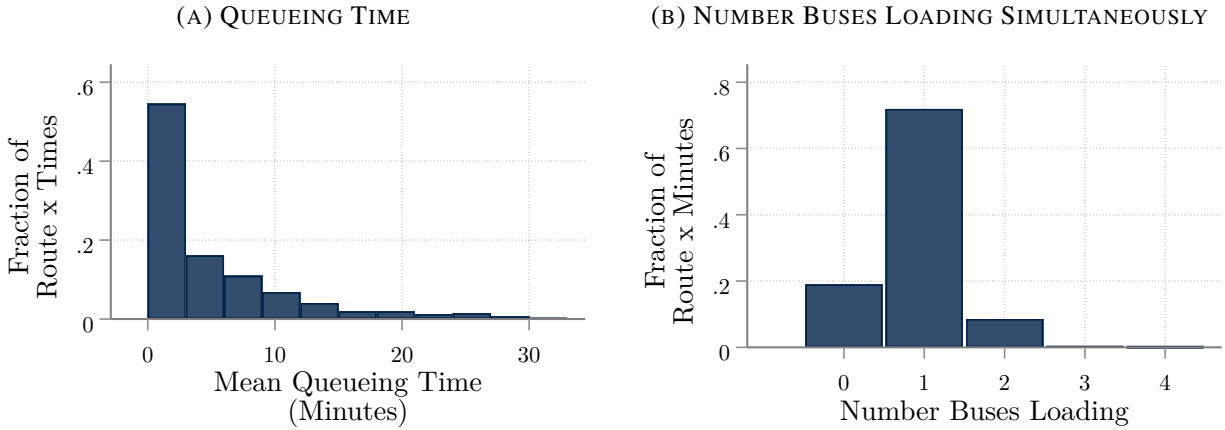
- Rose, John M., and Michiel C.J. Bliemer. 2009. “Constructing efficient stated choice experimental designs.” *Transport Reviews* 29 (5): 587–617.
- Schalekamp, H., and J. M. Klopp. 2018. “Beyond BRT: Innovation in Minibus-Taxi Reform in South African Cities.”
- Schalekamp, Herrie, and Nino McLachlan. 2016. “Minibus-taxi operator reforms, engagement and attitudes in Cape Town.” In *Paratransit in African Cities: Operations, Regulation and Reform*, edited by Roger Behrens, Dorothy McCormick, and David Mfinanga, 174–198. Routledge.
- Small, Kenneth A., and Erik T. Verhoef. 2007. *The Economics of Urban Transportation*. New York: Routledge.
- Spence, A. Michael. 1975. “Monopoly, Quality, and Regulation.” *The Bell Journal of Economics* 6 (2): 417–429.
- Springleer, C., A. Mulla, M. Moody, K. Kwinana, and C. Paulsen. 2023. “New Minibus-Taxi Initiatives in the City of Cape Town.”
- 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/conservation/public\\_transportation.html](https://afdc.energy.gov/conservation/public_transportation.html).
- Vitali, Anna. 2024. “Consumer Search and Firm Location: Theory and Evidence from the Garment Sector in Uganda.”
- Whittington, Dale. 2010. “What Have We Learned from 20 Years of Stated Preference Research in Less-Developed Countries?” *Annual Review of Economics* 2:209–236.
- Woolf, S.E., and J.W. Joubert. 2013. “A people-centred view on paratransit in South Africa.” *Cities* 35:284–293.
- Yürükoğlu, Ali. 2022. “Empirical Models of Bargaining with Externalities in IO and Trade.” In *Bargaining*, edited by Emin Karagözoğlu and Kyle B. Hyndman, 227–247. Palgrave Macmillan, Cham.
- Zarate, Roman David. 2024. “Spatial Misallocation, Informality, and Transit Improvements: Evidence from Mexico City.” *Revise & Resubmit, American Economic Review*.



## ONLINE APPENDIX

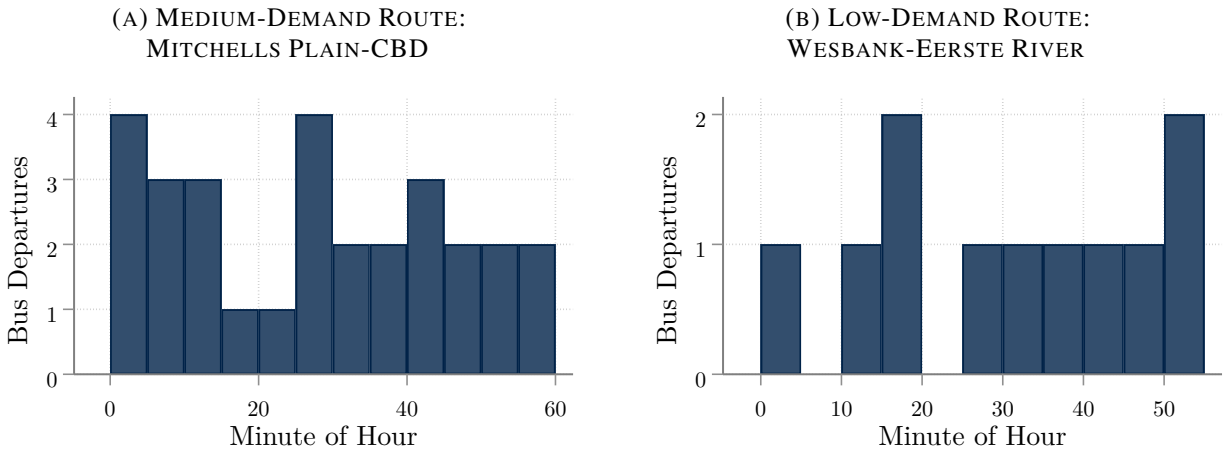
### A. ADDITIONAL FIGURES AND TABLES

**FIGURE A.1.** MINIBUS LOADING PROCESS: DISTRIBUTIONS OF . . .



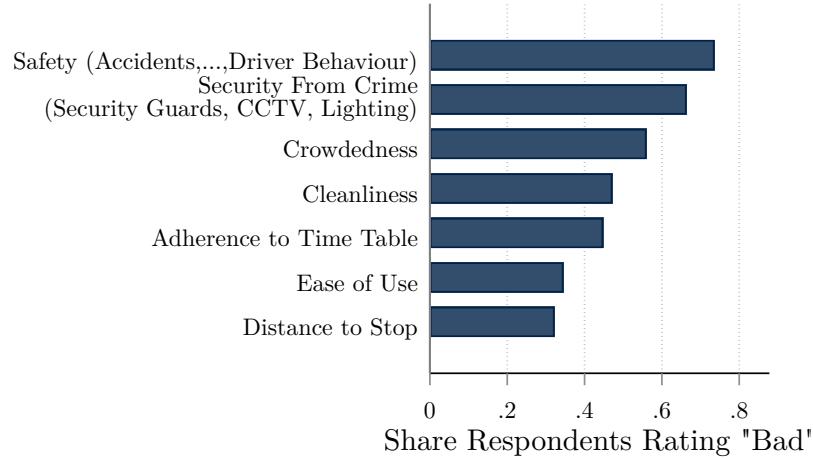
*Notes:* Panel (A) displays the distribution of passenger queuing times at the origin station for a specific route, over minibus routes and five-minute periods in my station count data from Cape Town. Panel (B) displays the distribution of the number of minibuses loading simultaneously at the origin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

**FIGURE A.2.** BUS DEPARTURE TIMES



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

**FIGURE A.3. RATINGS OF MINIBUS ATTRIBUTES**



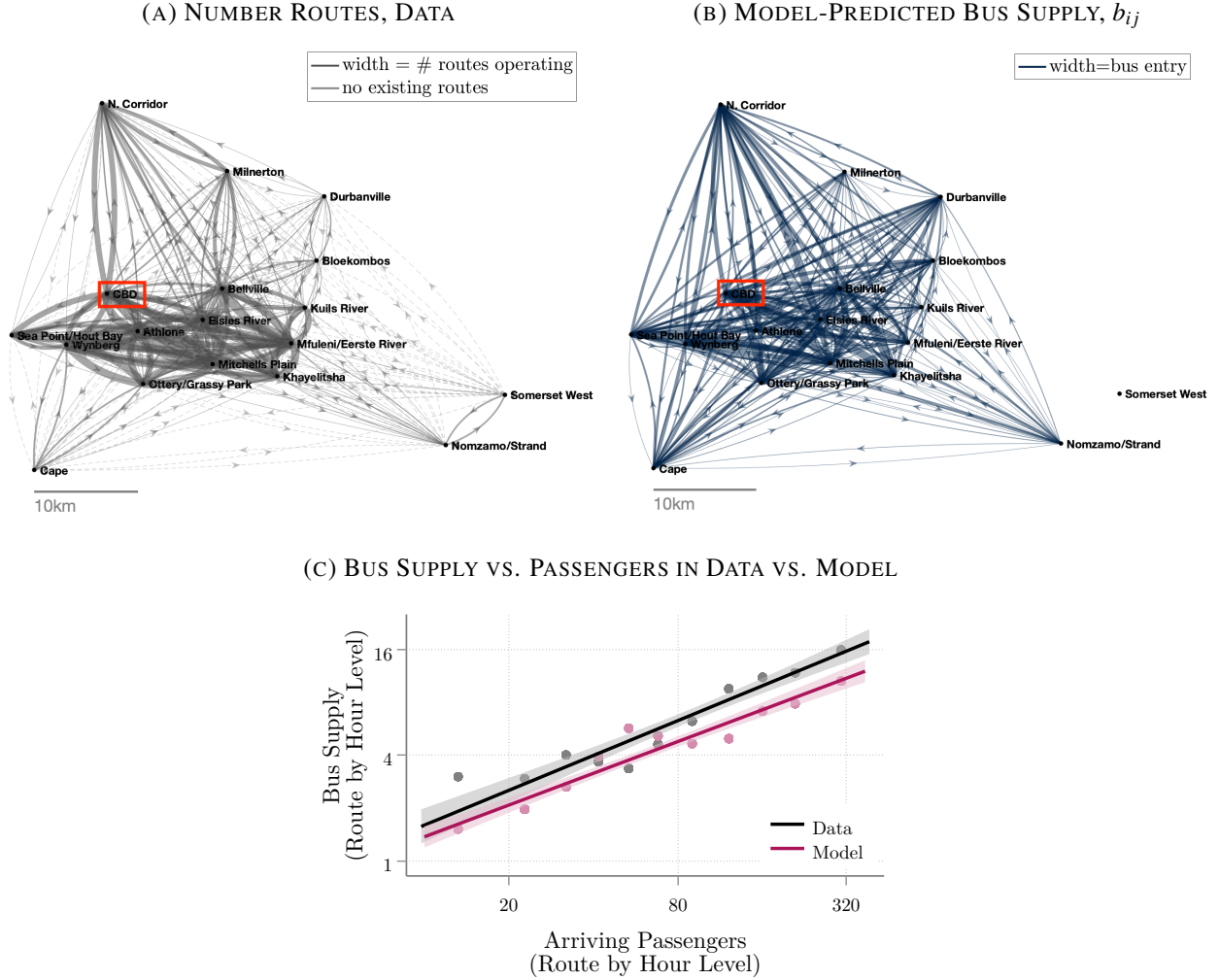
*Notes:* Figure displays a bar graph of the shares of respondents ( $N = 1685$ ) in the 2013 Cape Town Household Travel Survey rating each minibus attribute as “bad,” as opposed to “acceptable” or “good.”

**TABLE A.1. FORMALIZATION: DECOMPOSITION, MARKET MOHRING VS. BUS SUPPLY INSURANCE**

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
Full Formalization		0.05	0.04	-0.01	-0.02	0.45	-3.67	0.30	0.13	0.35	0.16
Fares Only		0.05	0.04	-0.01	-0.02	0.46	-3.6	0.27	0.1	0.33	0.13
Subsidies Only		0.001	0.001	-0.0001	-0.0007	-0.01	-0.07	0.03	0.03	0.03	0.03

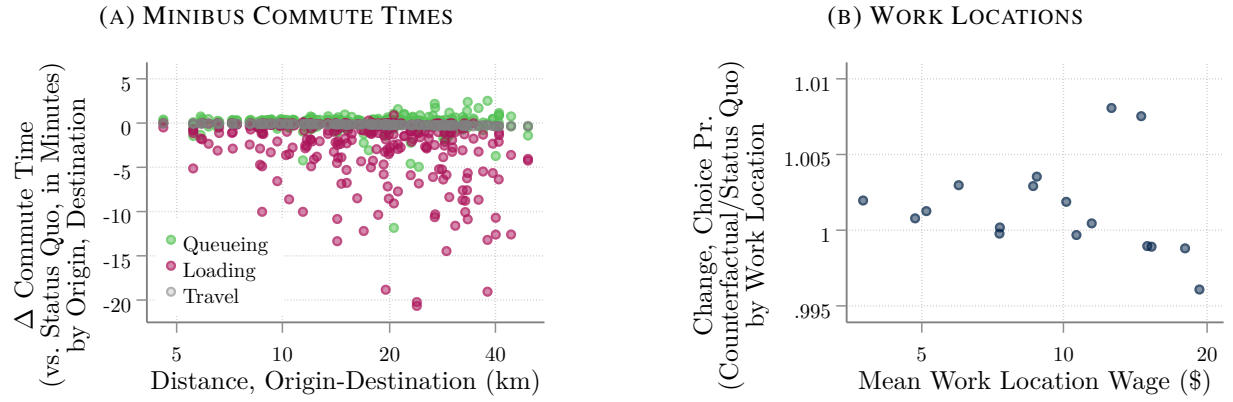
*Notes:* Table summarizes the effects of the full minibus formalization program of optimal fares and association subsidies discussed in the main text, as well as (i) a “fares only” program which imposes the formalization fare scheme, with association subsidies solely to close the gap between these lower optimal and the status-quo fares; and (ii) a “subsidies only” program which imposes the formalization scheme’s per-passenger association subsidies, with status-quo fares charged to passengers. 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.4.8. 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. Note that all changes are taken relative to the status quo.

**FIGURE A.4. MINIBUS NETWORK IN DATA VERSUS MODEL**



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

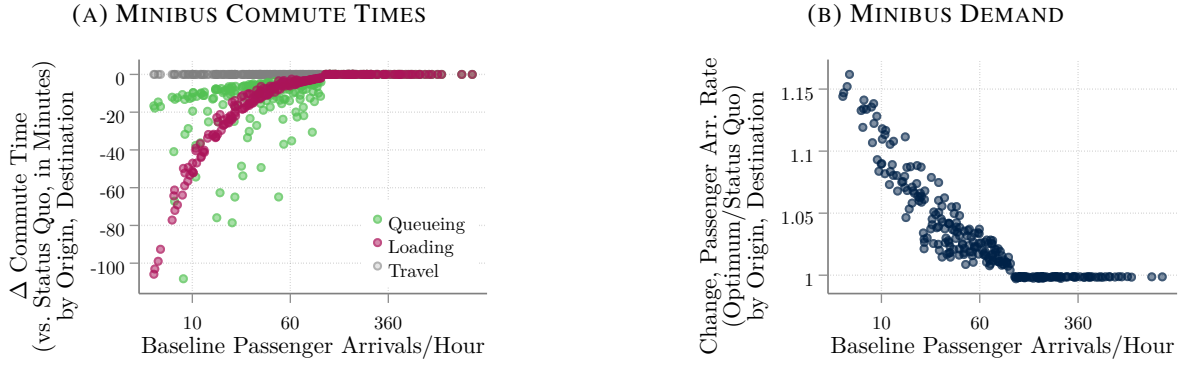
**FIGURE A.5. MINIBUS SPEED LIMIT ENFORCEMENT**



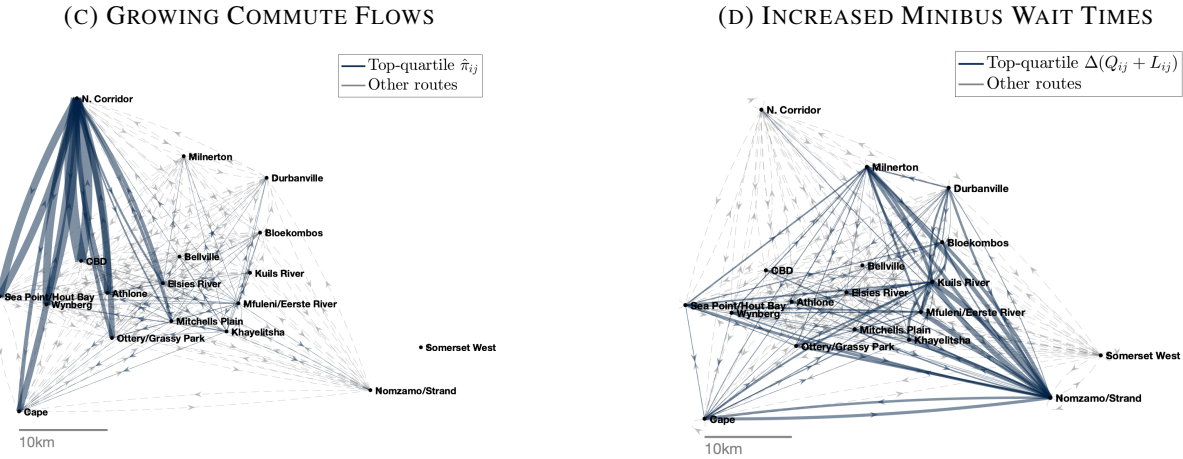
*Notes:* Figure characterizes speed limit enforcement counterfactual. Panel (A) displays, on the horizontal axis, the straight-line distance from a minibus route's origin to destination, and on the vertical axis, a scatterplot of route-level raw changes in queueing, loading, and travel times,  $Q_{ij}$ ,  $L_{ij}$ , and  $T_{ij}$ . Panel (B) displays a scatterplot of the proportionate change in work location choice probability, averaged over skill groups, versus the location's wage, again calculated as the average across skill groups, weighted by aggregate populations. Changes are calculated from the status quo to the speed limit enforcement counterfactual.

**FIGURE A.6. ALTERNATIVE POLICIES**

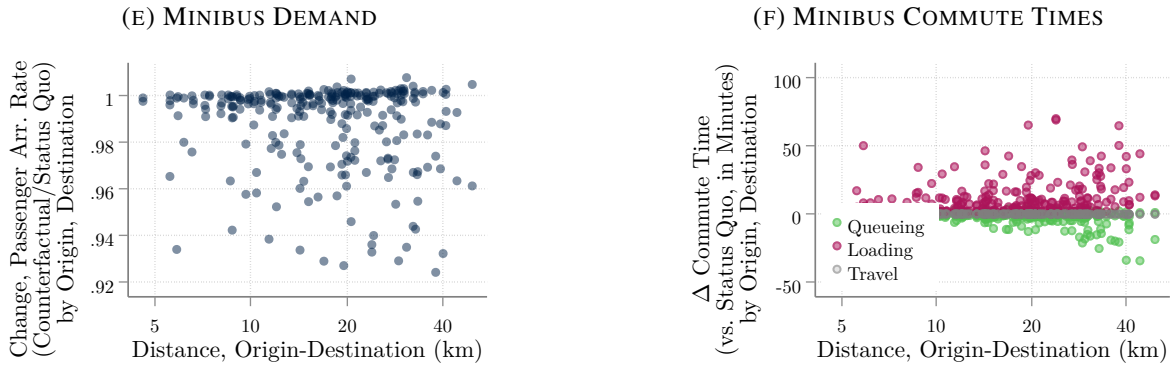
**MINIBUS SCHEDULE**



**MYCITI BUS RAPID TRANSIT (BRT)**



**LARGER MINIBUS “RECAPITALIZATION”**



*Notes:* Panel (A) displays, on the horizontal axis, the baseline route-level passenger arrival rate  $\lambda_{ij}$ , and, on the vertical axis, a scatterplot of route-level raw changes in queueing, loading, and travel times,  $Q_{ij}$ ,  $L_{ij}$ , and  $T_{ij}$  from the status quo to the schedule counterfactual. Panel (B)’s vertical axis instead displays a scatterplot of the changes in minibus passenger arrival rates under the schedule, relative to the status quo. Panel (C) maps the 25% of home-work location pairs with the largest proportionate increases in choice probabilities, from the counterfactual without MyCiti BRT to the status quo, where line width corresponds to the percent increase. Panel (D) instead maps the 25% of minibus routes with the largest raw increases in total wait times,  $Q_{ij} + L_{ij}$ , from the counterfactual without MyCiti BRT to the status quo, where line width corresponds to the relative magnitude of the increase. Panel (E) displays, on the horizontal axis, the straight-line distance from a minibus route’s origin to destination, and on the vertical axis, a scatterplot of minibus passenger arrival rates under the larger minibus recapitalization program, in changes relative to the status quo at the route level. Panel (F)’s vertical axis instead displays a scatterplot of route-level raw changes in queueing, loading, and travel times,  $Q_{ij}$ ,  $L_{ij}$ , and  $T_{ij}$ , from the status quo to the larger minibus counterfactual.

## B. DATA

### B.1 Minibus Station Counts

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

#### B.1.1 Sample

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

**Sampling Frame** Complicating the sampling procedure is the fact that no fully comprehensive, accurate list of stations and routes exists. Thus, as a sampling frame, I employ a roster of routes by origin station derived from a 2018 collaboration between GoMetro and the City of Cape Town’s Transport and Urban Development Authority.<sup>50</sup> This listing is, according to stakeholders, as comprehensive and accurate as any available and includes the number of minibus trips mapped by route in this previous data collection effort. Stakeholder discussions revealed that the number of trips mapped is an indicator of the number of buses on a route.

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

**Clusters and Stratification** I employed a two-stage stratified cluster sample, sampling 8 stations and then 6 routes originating at each of the 8 stations. I stratify within each stage by a proxy for station and route-level bus 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.

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.<sup>51</sup> Second, I sampled 3 stations, or a 16%

---

<sup>50</sup>In order to update the city’s record of on-the-ground minibus route paths and operations, GoMetro sent enumerators armed with a smartphone app to ride on close to 30,000 minibus trips on the approximately 800 established minibus routes. The results showed the official, city-designated routes to be outdated. For example, 250 of the official city routes no longer operated (Coetzee et al. (2018)).

<sup>51</sup>From the full listing of 31 stations, some which would have otherwise counted among the 5 busiest had to be skipped due to minibus associations denying permission (Nyanga Central, Gugulethu Eyona) or not containing 6 routes that load off-road (Claremont Station). The 5 busiest feasible stations are numbers 1-2, 4, 6, and 8.

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.<sup>52</sup>

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.

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

I now endeavor to compare the characteristics of routes in my sample with those of the entire population of routes in Cape Town. Note that I can do so only for 41 of my 44 sampled routes. Three routes in my sample, which were randomly chosen to replace infeasible or nonexistent originally-sampled routes, do not appear in the original sampling frame, so I lack data for these routes on the covariates which would allow comparison to the population of routes. For ease of interpretation, I now merge both directions of routes as well as any duplicate routes with the same origin and destination to obtain a population of approximately 429 unique routes in Cape Town.<sup>53</sup>

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

---

<sup>52</sup>Wynberg Station (Western Side) had to be excluded from the less busy sampling frame due to lack of permission, Cape Town CBD station due to lack of AM peak operations, and Mitchell’s Plain Station (North) and Promenade as well as Mitchells Plain Station Eastern Side (South) due to being adjacent to an already sampled station. After drawing the sample, further stations had to be excluded and resampled as follows. One station sampled, Khayelitsha (Vuyani), turned out to be part of an already sampled station (Khayelitsha (Nolungile Site C)); several others (Athlone, Vasco Station) do not operate as minibuses with queues and loading off-road, and another set (Zevenwacht Mall, Mitchells Plain (Promenade), Tableview (Bayside)) do not operate in the AM peak.

<sup>53</sup>Note that, unlike in my sampling frame, I do not exclude from this comparison population the stations with fewer than 6 routes.



**TABLE B.1. STATION COUNT SAMPLE CHARACTERISTICS**

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

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

### B.1.2 Data Collected

Station counts occurred on weekday mornings during one morning peak period (6-10am) per station, on weekdays from June 20-28 and 30, 2022. Two enumerators recorded data on each of 6 sampled routes over the course of the four-hour period. One enumerator stationed at the beginning of the loading lane and passenger queue corresponding to a route recorded, on forms as in Online Appendix Figure B.7a, 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.<sup>54</sup> The second enumerator per route monitored bus loading and departures, recording the time a vehicle pulled in to the front of the loading lane where buses typically load passengers, the time of departure, as well as the number of passengers on board at departure, as in Online Appendix Figure B.7b.

### B.1.3 Calculations

First, I calculate quantities related to the minibus loading process for the facts in Section III. I begin by recording the number of buses at the front of the loading bay, where passengers typically board, for each route and minute. I then discretize time into 5-minute periods, each beginning at some clock time  $t$ . I measure loading time  $L_{sijt}$  as the number of minutes between the time the minibus for trip  $s$  on route  $ij$  departing during time interval  $t$  arrived to the front of the loading bay and the time it departs, as described in Online Appendix Section B.1.2. I assume that the passengers I observe departing from the origin station on trip  $s$ ,  $deppax_{sijt}$ , board the bus at a uniform rate. Then, I proportionately apportion these passengers who depart on a trip to the five-minute blocks during which the bus was loading to calculate the total number of passengers

<sup>54</sup>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.

**FIGURE B.7. STATION COUNT DATA COLLECTION FORMS**

(A) PASSENGER QUEUES

GoMetro 40036 Yale Cape Town Surveys Rank Count Form | Passenger Waiting and Vehicle Arrival

Enumerator:	Rank:	Route:	Date:
Time	People Waiting	Time	People Waiting
06:00 – 06:05		06:10 – 06:15	
06:20 – 06:25		06:30 – 06:35	
06:40 – 06:45		06:50 – 06:55	
07:00 – 07:05		07:10 – 07:15	
07:20 – 07:25		07:30 – 07:35	
07:40 – 07:45		07:50 – 07:55	
08:00 – 08:05		08:10 – 08:15	
08:20 – 08:25		08:30 – 08:35	
08:40 – 08:45		08:50 – 08:55	
09:00 – 09:05		09:10 – 09:15	
09:20 – 09:25		09:30 – 09:35	
09:40 – 09:45		09:50 – 09:55	
Vehicle ID	Arrival Time	Vehicle ID	Arrival Time

(B) BUS LOADING AND DEPARTURE

GoMetro 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

*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.

boarding buses on route  $ij$  from  $t$  to  $t + 5$  as

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

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

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

under the assumption that no passengers abandon the queue after joining. I display these two variables,  $n_{ijt}$  and  $\lambda_{ij,t-1}$ , in Figure 6a.

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

Second, I detail the quantities needed to estimate queueing efficiency  $\mu$ , as in Equation (14). I already have bus-departure-level loading times  $L_{sijt}$ , which, indexed by the hour  $h$  and date  $d$  within which a 5-minute block  $t$  falls, are the  $L_{sijhd}$  on the left-hand side of (14). At the route  $ij$  by hour  $h$  by date  $d$  level, I calculate the probability  $p_{ijhd}^q$  of the existence of a (nonzero) queue of passengers, conditional on a bus being present, or “queue prevalence” for short. Concretely, this probability equals the share of every-5-minute observations snapshots, indexed by  $t$ , where I see one or more passengers in the queue ( $n_{ijt} > 0$ ) and a bus present in the loading area divided by the total share of these instants  $t$  with a bus present.

Third, I calculate variables for estimation of the minibus arrival efficiency  $\rho$  and internal calibration. To estimate arrival efficiency in (15), having already calculated  $b_{ijhd}$  and  $L_{sijhd}$ , I calculate, for minibus trip (i.e. departing bus)  $s$ , the minutes elapsed between the departure of the previous bus and the instant when bus  $s$  arrives to the front of the loading bay, which yields  $\text{gap}_{sijhd}$ . Using my separate onboard tracking data, I calculate the average origin-destination travel time on route  $ij$ , which yields  $T_{ij}$ . For internal calibration, I calculate the median minibus fare by first taking a passenger-weighted average by route in the same onboard tracking data and then taking the median across routes. The median queue length in the internal calibration equals the median of  $n_{ijt}$ .

## B.2 Minibus Stated Preference Survey

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

### B.2.1 Questionnaire Design

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

**Choice of Attributes and Levels** In a discrete choice experiment, attributes should be chosen that are important and relevant to the decision at hand (Mangham et al. (2009) and Johnston et al. (2017)). Conveniently, the aforementioned Cape Town Household Travel Survey asks respondents to rate the importance of a variety of factors in their mode choice decisions; the three factors most frequently rated “most important” in mode choice were comfort, safety, and security. In separate questions asking respondents to rate various aspects of existing minibus service on a four-point scale, “Safety (accidents, maintenance, driver behavior),” “Security from crime,” and “Availability of a seat/crowdedness” are also those most frequently rated “bad.”<sup>55</sup>

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)),

---

<sup>55</sup>Other aspects included timetable adherence, cleanliness, distance to stop, and ease of use.

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.<sup>56</sup>

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

**Coefficient Priors** To implement the d-efficiency algorithm and thereby choose attribute levels for the final survey, I now require priors for the demand model parameters. I obtain priors  $v = 12.7$ ,  $r = 0.002$ , and  $\kappa_M = 0.88$  from estimating a mode choice model on the Cape Town household travel survey stated preference module.<sup>57</sup> Then, since no larger-sample stated preference surveys cover analogous attributes, I use the results of my 1-day stated preference pilot survey—which surveyed  $N = 20$  commuters using a very similar format—to estimate priors for  $\xi_z$ ,  $z \in \{\text{security, no speeding, no overloading}\}$ , obtaining  $\xi_{\text{security}} = -0.18$ ,  $\xi_{\text{no speeding}} = -0.16$ , and  $\xi_{\text{no overloading}} = -0.21$ .<sup>58</sup> 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.<sup>59</sup> 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.

---

<sup>56</sup>Fares can take values of R6, R10, R14, and R18, while travel time is either 20, 30, 40, or 50 min., corresponding to the lengths of typical minibus rides in the morning peak.

<sup>57</sup>I restrict the sample to choice sets that do not contain car as a mode and to respondents aged 25-65 who work outside the home.

<sup>58</sup>In estimating the multinomial logit on pilot data, I restrict the coefficients on travel time, cost, and the travel-time income interaction to be consistent with the aforementioned two priors and use the midpoints of household income bins from a separate income question.

<sup>59</sup>I create two blocks, which are randomized across respondents, because doing so increased power in Monte Carlo simulations without increasing respondent burden. As for the numbers of choice sets and alternatives, I reduced these from 8 to 5 and 3 to 2, respectively, after the pilot revealed respondent frustration and inattention towards the end of the survey – and a version with 2 rather than 3 alternatives proved less problematic in this regard.

### *B.2.2 Pilot Survey Lessons*

The enumerator team and I conducted a pilot survey at the Cape Town CBD minibus station on 15 June 2022 from approximately 11am to 1pm, where we contacted 36 respondents, 25 of whom qualified for and completed the pilot questionnaire, which had one version (block) with 9 choice sets of 3 alternatives each and a second block with 9 choice sets of 2 alternatives each. Anecdotally, the respondents I interviewed seemed to be taking the scenarios seriously and understanding the aim of the exercise, musing out loud, for example, “I can’t take this bus because it will make me late to work!” However, after the pilot survey, I reduced the 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 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.<sup>60</sup>

### *B.2.4 Administration and Script*

Survey enumerators randomly approached respondents and asked for their consent to participate in the survey, according to a script in Online Appendix Figure B.8a. 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.8b-B.8d), 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.

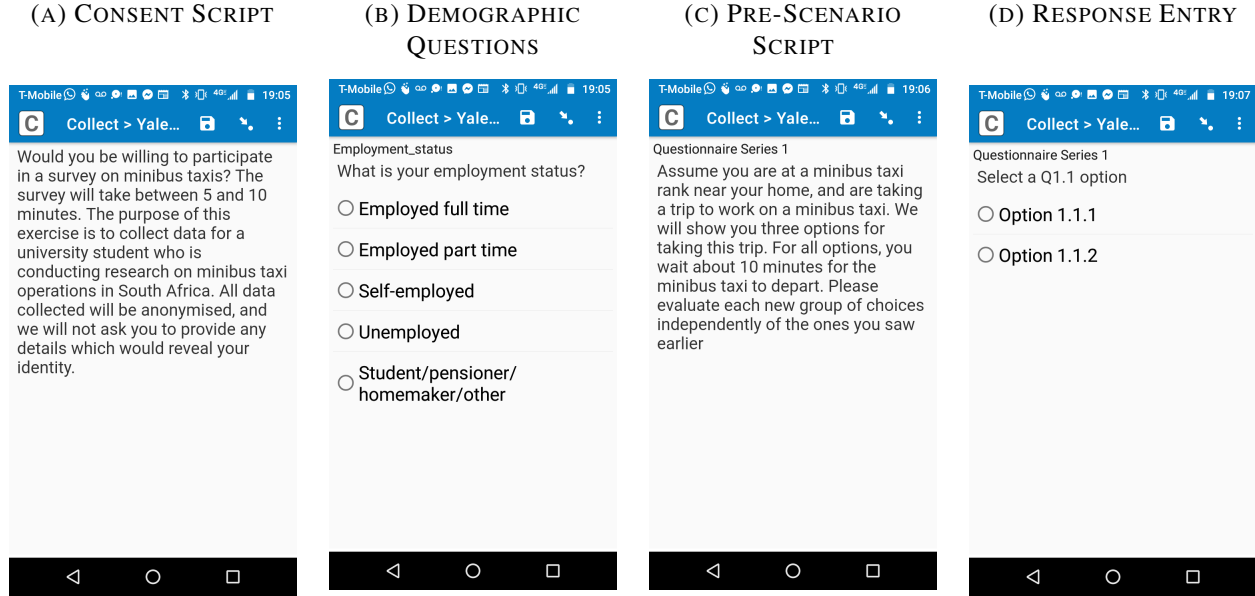
### *B.2.5 Field Experience*

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.

---

<sup>60</sup>Enumerators approached 586 people. Of these, 333 were full-time employed, 97 part-time employed, and 96 self-employed, for a total of 526 respondents who qualified for the survey. Of the remaining people not qualifying, 14 were students/pensioners/homemakers/other and 42 were unemployed.

**FIGURE B.8. 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.2 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)

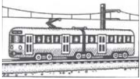


The City of Cape Town conducted the 2013 Cape Town Household Travel Survey (CTHHTS) on a representative sample of residents. In addition to demographics and car ownership, this survey records the addresses of residence and work, which I successfully geocode with the Google Geocoding API for  $N = 17,395$  employed respondents, along with the details of respondents' commutes. For descriptive statistics and moments, I define the commute mode as follows: minibus 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.

**TABLE B.2. STATED PREFERENCE SAMPLE CHARACTERISTICS**

Variable	<i>Stated Pref. Samples</i>		<i>Data</i>
	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	
<i>N</i>	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.

**FIGURE B.9. EXAMPLE CITY OF CAPE TOWN STATED PREFERENCE SURVEY SCENARIO**

Choice 1:	Train <input type="checkbox"/>	MyCiti Bus <input type="checkbox"/>	Minibus Taxi <input type="checkbox"/>
Cost	R 3	R 20	R 20
In-Vehicle Travel Time	70 min	25 min	70 min
Waiting time	15 min	20 min	10 min
Number of Transfers	2	2	2
No Preference <input type="checkbox"/>			

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

A subset of respondents also completed a commute stated preference survey with a format similar to my own, except that respondents chose among different *modes* of transport. Respondents were asked to consider their last commute to work or education and then indicate their preferred mode option from each of 13 choice sets, of which Online Appendix Figure B.9 displays an example. Each choice set offered 3 mode options from among car, formal MyCiti bus, (regular) formal bus, formal train, or minibus (taxi). Each option in a given choice set, in addition to representing a different mode, varied in monetary cost, travel time, waiting time, and number of transfers. However, this city-run survey did not include non-pecuniary quality improvements such as station security.



## B.4 Minibus On-Board Tracking Data

I make use of GPS-tracked minibus trips, also newly collected for this paper by the South African firm GoMetro. This data, logged by enumerators via a smartphone app, covers two trips from the beginning to the end of each route in my station count data and provides stop-level information within each trip. For each stop, I observe the number of passengers boarding and alighting, the arrival and departure time, and the fare paid by passengers boarding. In total, my sample includes  $N = 582$  stops, made by 60 vehicles on 43 routes over 2 trips per route.

## C. THEORY

### C.1 Micro-Foundations of Commute Utility

I now lay out a micro-founded model of commuting which underlies the linear approximations to commute utility  $U_{ijm}^g$  used in the main text. Suppose that commuters, upon birth, immediately enjoy their home-location-specific amenity  $\theta_i^g$ , pay a one-time mode-specific utility cost  $\kappa_M^g$  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  $\tau_{ijm}$ . Given the rate of time preference  $r$ , the value of waiting,  $u_{ijm}^g$ , satisfies

$$u_{ijm}^g = \exp(-rw_{ijm}) (V_{ijm}^g - \tau_{ijm}), \quad (\text{C.1})$$

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  $\omega_j^g$  received upon arrival, discounted to account for travel time  $t_{ijm}$ :

$$V_{ijm}^g = \exp(-rt_{ijm}) \omega_j^g, \quad (\text{C.2})$$

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

$$u_{ijm}^g \approx \omega_j^g - \tau_{ijm} - r\omega_j^g(w_{ijm} + t_{ijm}) + r\tau_{ijm}w_{ijm} \quad (\text{C.3})$$

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

$$\bar{U}_{ijm}^g \approx \theta_i^g - \underbrace{\kappa_m^g - r\omega_j^g[E(w_{ijm}) + t_{ijm}] - \tau_{ijm}}_{\equiv U_{ijm}^g} + \omega_j^g. \quad (\text{C.4})$$

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 (8) 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  $\tau_{ijm} = \tau_A$ .

## C.2 Empirical Queueing Model

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

$$\frac{1}{\lambda_{ij}^B} \equiv \frac{\log \left\{ \exp \left[ \rho \left( \frac{2T_{ij}}{b_{ijhd}} - L_{nijhd} \right) \right] + 1 \right\} - \log \varsigma_{ij}}{\rho}. \quad (\text{C.5})$$

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

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

$$L_{ij} = \frac{1}{\iota_{ij}} + \frac{1}{\mu_{ij}^B} = \frac{\bar{\eta}}{\mu p_{ij}^{q|b}} + \underbrace{\frac{1}{\iota_{ij}}}_{\equiv \varepsilon_{ij}}, \quad (\text{C.6})$$

where I have used the expression for  $\mu_{ij}^B$  from the main text and define the idiosyncratic error  $\varepsilon_{ij} \equiv \frac{1}{\iota_{ij}}$ . Rewriting this equation at the individual bus departure level yields my main specification (14) for  $\mu$ . In turn, for bus arrival efficiency  $\rho$ , my specification (15) is simply a bus-departure-level version of (C.5).

## D. ESTIMATION

### D.1 Queueing Efficiency

#### D.1.1 First Stage Specifications

In this section, I present the first-stage regressions from 2SLS estimation of Equation (14) in Table D.1. Table 2 in the main text, in turn, contains the second-stage results.

#### D.1.2 Threats to Identification

I now probe the plausibility of the exclusion restriction for the queueing efficiency regression: namely, that idiosyncratic interruptions to loading do not vary systematically with work start times. Recall that the

**TABLE D.1. QUEUEING EFFICIENCY: FIRST STAGE**

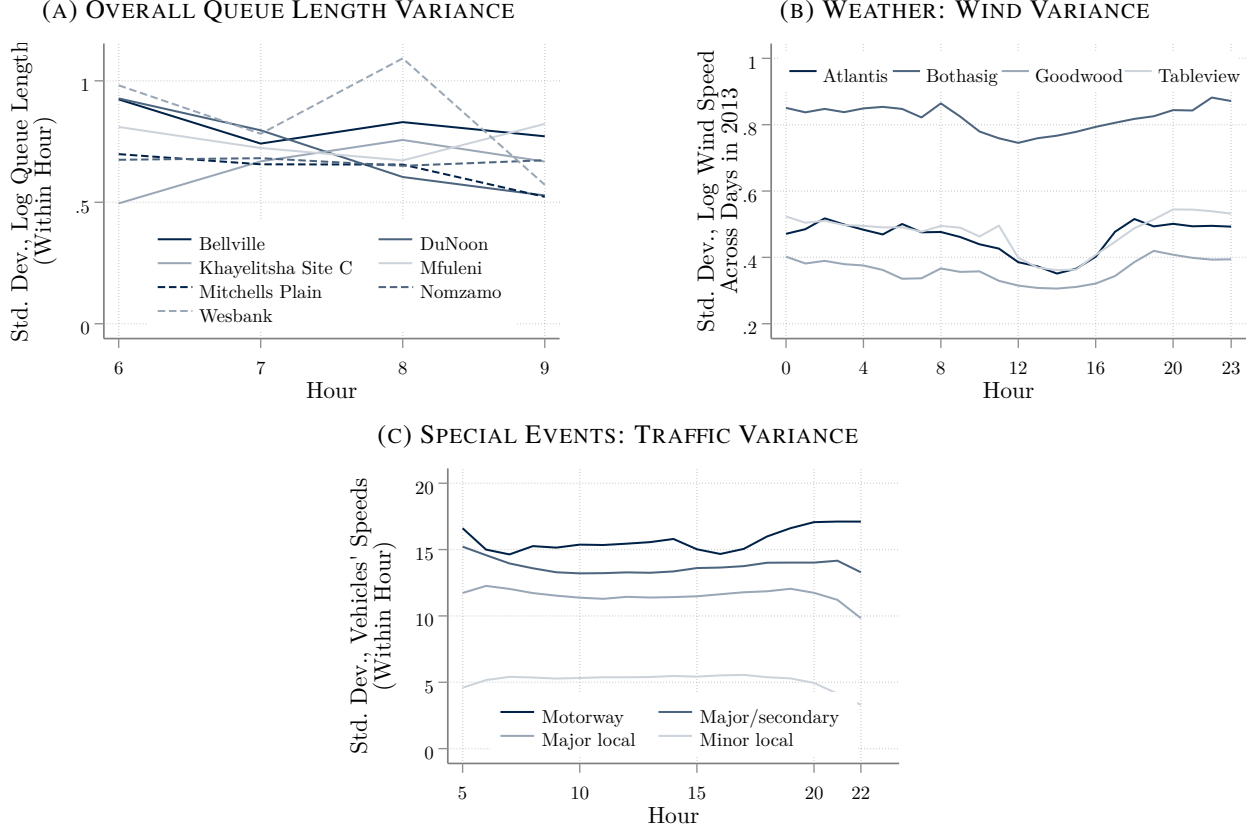
Parameter	(1) $\left(p_{ijhd}^{q b}\right)^{-1}$	(2) $\left(p_{ijhd}^{q b}\right)^{-1}$	(3) $\left(p_{ijhd}^{q b}\right)^{-1}$
log number commuters	-0.41 (0.17)	-0.43 (0.12)	-0.63 (0.33)
Route-by-Date FE		✓	✓
Origin-by-Hour-by-Date FE			✓
Observations	1101	1101	1101
F Statistic	6.16	12.21	3.69

*Notes:* Robust standard errors in parentheses, clustered at the route-by-hour-by-date level. Table presents the first-stage specifications from 2SLS estimation of (14) over bus departures in my station count data. The first stage regresses the (inverse) probability of a nonzero passenger queue conditional on a bus being present in the loading area on the log number of commuters living in the mesozone spatial unit where the route originates and working in the route’s destination mesozone who report leaving their home during hour  $h$ , calculated from the 2013 Cape Town Household Travel Survey, along with fixed effects, as noted, for route-by-date and origin-by-hour-by-date. I report the effective first-stage F statistic following Olea and Pflueger (2013).

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 throughout the 6-10am morning commute. In sum, available evidence suggests that key sources of minibus loading delays vary little over the course of the morning rush hour, so any correlation with my work start times instrument must be mild.

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



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

## D.2 Minibus Arrival Efficiency: Details

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

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

across bus departures  $s$  on route  $ij$  from neighborhood (Transport Analysis Zone, i.e. TAZ)  $n(i)$  to neighborhood  $n(j)$  during hour  $h$  on date  $d$  in my station count data. The unobserved arrivals efficiency,  $\zeta_{sijhd}$ , reflects a variety of factors outside the model likely to impact associations' bus supply: overall foot and car traffic levels in origin neighborhoods as well as weather, special events, and road closures. I neutralize traffic and weather around the origin minibus station with origin neighborhood-by-hour-by-date fixed effects and delays

**TABLE D.2. MINIBUS ARRIVAL EFFICIENCY ESTIMATES**

	<i>GMM-IV</i>	
	(1) bus gap	(2) bus gap
$\rho$ <i>Arrival Efficiency</i>	0.177 (0.09)	0.178 (0.09)
Origin Neighborhood-by-Date FE	✓	
Destination Neighborhood FE	✓	✓
Hour FE	✓	
Origin Neighborhood-by-Hour-by-Date FE		✓
Observations	1,158	1,158

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

common to destinations with destination neighborhood fixed effects. To address any differential impacts of, say, foot traffic on different routes operating out of the same origin neighborhood, I instrument for the first term on the right-hand side, which varies primarily with bus supply  $b_{ijhd}$  on a route, with the distance from that route’s origin to destination. Route length determines operations costs and thus supply but should not systematically vary with the degree to which, say, local traffic or road closures affect specific routes at the same station.

In my baseline specification in Column (1) of Table D.2, I find a bus arrival efficiency of  $\hat{\rho} = 0.177$ . When I add origin neighborhood-by-hour-by-date fixed effects in Column (2), the estimate remains almost unchanged; I employ the Column (2) estimate in the model quantification and counterfactuals in Sections VI-VII.

### D.3 Stated Preference: Sample Robustness

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.3 shows that the estimated parameters, though somewhat noisier, mirror my full-sample estimates quite closely, in particular the rate of time preference  $r$ , Gumbel scale  $v$ , 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 the 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 takeaways remain.

**TABLE D.3. 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)
$\kappa_M$	<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)
$\xi_{\text{security}}$	<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)
$\kappa_F$	<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)
$N$ 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.

## D.4 Externally Calibrated Parameters

### D.4.1 Geography

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

#### D.4.2 Commuter Populations $N^g$ and Average Wages

I calibrate commuter populations  $N^g$  using the 2013 Cape Town Household Travel Survey (see Online Appendix B.3) and the accompanying sample weights. Since commuters in my model cannot work in their home zone, I exclude those who work in the same transport analysis zone, as well as the less than 5,000 residents of the aforementioned one primarily non-residential TAZ, from these commuter populations. Since my model and data apply to the 6-10am peak-hour commute, I rescale these populations by the share (84%) of Cape Town commuters who start work within these four hours and then divide by 240 to calculate  $N^g$  as a per-minute inflow. I define two skill groups  $g$ , high and low, where high-skill includes those with a tertiary degree. For the stated preference estimation and normalization of model wages, I also use the 2013 Cape Town survey to compute average wages by skill group, taking the weighted mean daily per-person household income of workers in a skill group employed outside the home.<sup>61</sup>

#### D.4.3 Road Congestion Elasticity $\gamma$

Table D.4 displays the results from a regression, in TomTom MOVE API data for Cape Town, of log segment travel time on traffic volume, additionally interacted with road type indicators. I further discuss the specification in the main text in Section V.

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

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

#### D.4.5 Formal Transit Wait and Travel Times $H_{ij}, T_{ijF}$

I calibrate the full origin-destination matrix of formal transit wait,  $H_{ij}$ , and travel times,  $T_{ijF}$ , using a stylized network. The formal transit network consists of links between every pair of locations, and the Microsoft Azure API provides average formal transit wait and travel times along each link, where, crucially, these calculations allow commuters to walk to transit stops and transfer, if necessary.<sup>62</sup> Then, for each origin-destination pair  $ij$ , I find the shortest path through the network, accounting for wait and travel time along each link; the respective sums yield the total wait time,  $H_{ij}$ , and the total travel time,  $T_{ijF}$ , even for origin-destination pairs where no direct connection exists.

---

<sup>61</sup>To calculate daily personal income  $\omega_i$  in the Cape Town survey, I take the midpoint of the household's income bin, divided by 22.5 (the number of working days in a month) times the number of people in the household. Additionally, I multiply by  $\frac{1}{2}$  since I only model a one-way commute. For my own survey, I make similar adjustments, except that I have personal income directly instead of needing to impute it from household income. Finally, I convert all monetary amounts, including fares, to USD for scaling purposes.

<sup>62</sup>Specifically, from the Microsoft Azure API (similar to Google Maps), I obtain formal transit wait and travel time for each link in the formal 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. Travel time includes any walking and transfer time.



**TABLE D.4.** ESTIMATION OF ROAD CONGESTION ELASTICITY  $\gamma$ 

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

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

#### D.4.6 Formal Transit Fares $\tau_{ijF}$ and Car Commute Cost $\tau_A$

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  $\tau_A$  using an “average [monthly] total mobility cost ” calculated by WesBank of South Africa.<sup>63</sup>

#### D.4.7 Minibus Operations Parameters $\bar{\eta}$ , $\chi_{ij}$

I set minibuss capacity to the  $\bar{\eta} = 15$  passengers in line with the size of 94% of minibuses, as discussed in Fact 3. I parameterize the minibuss operating cost matrix  $\chi_{ij}$  as proportional to driving distance and then use fuel efficiency figures provided by the firm GoMetro to calibrate the per-kilometer cost.<sup>64</sup>

<sup>63</sup>The monthly [average total mobility cost](#) equals ZAR 9,356.80; I divide this figure by the number of (half-)working days per month, and convert to USD.

<sup>64</sup>I use the driving distance under free-flow (Sunday, 11pm) conditions between the centers of employment of transport analysis zones (TAZ)  $i$  and  $j$  predicted by the Google Maps Distance Matrix API. For the per-kilometer cost, I multiply the Toyota Quantum minibuss’s litres of diesel used per kilometer, 0.099, by the June 2022 diesel per-litre price in South Africa, ZAR22.63, and convert to USD, in line with other prices in my model.

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

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

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

#### D.4.8 Emissions Parameters $\chi_m^e$ and $\varsigma$

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 CO<sub>2</sub>-equivalent emissions per kilometer from driving, a figure which includes actual CO<sub>2</sub> 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)).<sup>65</sup> For the social cost of carbon, I follow Borck (2019)’s benchmark of  $\varsigma = \$0.0485$  per kilogram CO<sub>2</sub>. The per-commuter emissions cost on mode  $m$  from home  $i$  to work  $j$  then equals  $E_{ijm} \equiv \varsigma \chi_m^e \times \text{driving distance}_{ij}$ , where I calculate driving distance as in Online Appendix D.4.7.

## 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 minibuses and cars, the model-predicted mode shares by origin-destination pair and skill are significantly positively correlated with their empirical counterparts.

<sup>65</sup>Fuel 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  $\chi_A^e$  yields  $\chi_M^e = 0.0615$ . The AFDC estimates 137.2pmpg for high-ridership buses and 600pmpg for high-ridership trains. 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.2.** STATED PREFERENCE HETEROGENEITY: DIFFERENCE IN PARAMETER ESTIMATE, VERSUS BASE CATEGORY

Dimension	$r$	<i>Mode Utility Cost</i>		<i>Effects on Minibus Utility Cost</i>		
		$\kappa_M$	$\kappa_F$	$ \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 Household Travel Survey stated preference modules.

## 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.

## E.3 Counterfactual Robustness

In this section, I show how alternative modeling assumptions alter the welfare gains from my primary counterfactual, namely the formalization program fares and subsidies simulated in Section VII. I re-solve for the baseline equilibrium under each set of alternative assumptions.

### *Nested Logit*

First, 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  $\varepsilon_{ij}$  for each home-work location pair, with variance scaled by the parameter  $\zeta$ , 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  $\nu$  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  $\bar{U}_{ij}^g \equiv \nu \log \left[ \sum_{m'} \exp\left(U_{ijm'}^g\right)^{1/\nu} \right]$  such that commuters then choose a given home-work pair with probability

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

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 (9) in the equilibrium definition and pin down choice probabilities  $\pi_{ijm}^g$ .

I set  $\nu = 4.76$  as in the main text and, given the lack of available estimates for South Africa, calibrate  $\zeta$  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  $\zeta = 5.74$ .<sup>66</sup>

This nested logit demand system, as calibrated, now means commuters less willingly change home or work locations than they do modes. In the baseline equilibrium, the minibus mode share is somewhat higher, so the gains from the formalization program, gross of emissions, in Online Appendix Table E.3 also turn out marginally higher. However, the geography of commute distances in Cape Town interacts with commuters' lower location choice elasticity to reduce the magnitude of emissions reductions, so the net-of-emissions welfare gains from formalization are slightly lower than in the baseline.

---

<sup>66</sup>Specifically, let  $\tilde{\zeta}$  given Frechet elasticity. Denote the expected utility of home-work pair  $ij$  for group  $g$  by  $\bar{\Omega}_{ij}^g \equiv \theta_i^g + \bar{U}_{ij}^g + \omega_j^g$ ; then, the Frechet elasticity satisfies  $\tilde{\zeta} = \frac{\partial \log(\pi_{ij}^g/\pi_{kl}^g)}{\partial \log \bar{\Omega}_{ij}^g}$ . In my logit demand system, instead,

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

where the final equality uses the chain rule. Thus, a parameter  $\zeta$  in my model equivalent to the Frechet parameter  $\tilde{\zeta}$  must satisfy  $\zeta = \frac{\bar{\Omega}_{ij}^g}{\tilde{\zeta}}$ . I calculate the mean realized value of  $\bar{\Omega}_{ij}^g$  in my baseline model, weighted by the commuter populations of each skill  $g$  commuting from  $i$  to  $j$ , and divide by  $\tilde{\zeta} = \eta_l = 2.07$  from Tsivanidis (2023) to obtain  $\zeta = 5.74$ .

### Non-Rush-Hour Demand

Second, I consider the ramifications of my formalization policy outside of rush hour. Approximately 84% of work commutes, as reported in the 2013 Cape Town Household Travel Survey, occur during the 6-10am rush hour, which I study and simulate in the main text. Outside of rush hour, non-work trips will make up a much more sizable share of total demand. I could capture non-work trips by incorporating a framework along the lines of Miyauchi et al. (2022); in this section, I provide an indication of how the welfare effects of formalization might change during non-peak hours by simulating the policy under commuter inflows  $N^g$  only half as high as in the baseline model. In the second row of Online Appendix Table E.3, I find even larger gains from formalization under this alternative, lower-demand scenario. Intuitively, due to the Market Mohring Effect, minibuss routes suffer even longer wait times when demand is only half as high, which increases the gains from growing the “scale” of minibuss commuting via optimal fares and subsidies.

### Endogenous Bus Departures

Third, I model and quantify the implications of associations’ choice of the number of passengers  $\eta_{ij}$  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*  $\eta_{ij}$ , a bus departs at rate  $\mu_{ij}^B \equiv \mu p_{ij}^{q|b} / \eta_{ij}$ , where  $p_{ij}^{q|b}$  denotes the probability that a nonzero queue of passengers is waiting, conditional on a bus being present in the loading area.

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

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

Associations’ first-order condition for the departure threshold,

$$0 = \frac{\partial \Pi_{ij}}{\partial \eta_{ij}} = \left( \tau_{ij} - \frac{\chi_{ij}}{\eta_{ij}} \right) \frac{\partial \lambda_{ij}}{\partial \eta_{ij}} + \frac{\lambda_{ij} \chi_{ij}}{\eta_{ij}^2}, \quad (\text{E.4})$$

highlights the tradeoffs at work in associations’ decisions 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 costs of a trip to be “spread” across more fare-paying commuters. For the sake of brevity, I avoid restating the equilibrium conditions and formal equilibrium definition. Instead, equilibrium now consists of a vector  $\{b, \tau, \pi, \eta\}$  that satisfies (i) (6), (7), and (9), with  $\eta_{ij}$  always in place of  $\bar{\eta}$ ; (ii) the departure threshold first-order condition (E.4); and (iii) the bus capacity constraint  $\eta_{ij} \leq \bar{\eta}$ .

I recalculate the effects of the formalization fares and subsidies from the main text in this new model. In the baseline equilibrium with endogenous departures, all routes depart full, i.e. with  $\eta_{ij} = \bar{\eta}$ . The higher gross-of-subsidy fares in the formalization program increase the marginal revenue of the additional passengers which leaving “early” attracts, so the share departing full falls to 83.9%. The resulting slightly shorter loading

**TABLE E.3. ROBUSTNESS: FORMALIZATION COUNTERFACTUAL**

<b>Model</b>	<i>Skill:</i>	Change in Mode Share				% Change in...					
		Minibus		Car		Earned Wage	Emissions	Welfare		Welfare, Net of Emissions	
		<i>Low</i>	<i>High</i>	<i>Low</i>	<i>High</i>			<i>Low</i>	<i>High</i>	<i>Low</i>	<i>High</i>
Nested Logit		0.05	0.04	-0.01	-0.02	0.44	-3.62	0.31	0.14	0.33	0.15
Non-Rush Hour		0.05	0.04	-0.01	-0.02	0.46	-3.71	0.32	0.15	0.37	0.17
Endogenous Bus Departures		0.05	0.04	-0.01	-0.02	0.45	-3.63	0.30	0.14	0.35	0.16
Agglomeration		0.05	0.04	-0.01	-0.02	0.53	-3.68	0.34	0.16	0.40	0.19

*Notes:* This table summarizes the effects of the formalization counterfactual, under four variations on the model in Section IV in the main text: (i) nested logit commuter demand, with a nest for each home and work location pair comprising the three modes; (ii) “non-rush hour” commuter demand half as high as in the baseline; (iii) associations choose with how many passengers, on average, to depart; and (iv) wages which increase with location employment, using the agglomeration elasticity in Ahlfeldt et al. (2015). The first four columns show the changes in the minibuss 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.4.8. 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.

times, however, do not affect the welfare gains in the third line of Online Appendix Table E.3.

#### *Agglomeration Spillovers*

Fourth, I augment the model with agglomeration spillovers. In particular, I let wages in a location increase with total local employment  $N_j \equiv \sum_{i,m,g} N^g_{ijm}$  according to  $w^g_j \equiv \xi^g_{0j} N^{\xi_1}_j$ . The equilibrium conditions are then identical to those in the model in the main text, except with the endogenous wages  $w^g_j$  in place of the exogenous  $\omega^g_j$ . Following Ahlfeldt et al. (2015), I set  $\xi_1 = 0.07$  and calibrate the  $\xi^g_{0j}$  such that  $w^g_j = \omega^g_j$  in the baseline equilibrium. For tractability, I assume that neither associations’ bus supply choices nor the Nash bargaining of fares account for the respective effects on wages  $w^g_j$ . In the fourth line of Online Appendix Table E.3, I find that accounting for agglomeration spillovers increases the gains from formalization, as workers concentrate in ex-ante higher-wage locations and thereby further push up wages.