Are There Too Many Minibuses in Cape Town? Privatized Provision of Public Transit*

Lucas Conwell †

October 29, 2022

Job Market Paper

[*Please* click here *for updated version*]

Abstract

Workers in low- and middle-income countries waste significant amounts of time commuting, partly due to gaps in public transit. In many African cities, privately-operated minibuses provide 50–100% of urban transit, at the cost of long wait times and poor personal safety for riders. Which externalities exist in the seemingly chaotic minibus market, and can policy interventions improve the market allocation? I build a micro-founded model of privatized shared transit subject to externalities in matching between buses and passengers, security provision, and road congestion. I then estimate the model with newly-collected data on minibus operations in Cape Town and stated user preferences for exogenously-varied commute attributes. I find that an optimal subsidy on minibus entry to correct matching externalities increases welfare and benefits low-skill workers on long routes. Government actions to improve security bring even more substantial welfare gains.

*I am deeply indebted to my advisors Costas Arkolakis, Michael Peters, Mushfiq Mobarak, and Orazio Attanasio for their invaluable guidance and generosity with their time and ideas. I additionally thank Lorenzo Caliendo, Richard Carson, Will Damron, Fabian Eckert, Cecilia Fieler, Claudia Gentile, Antonia Paredes-Haz, Sam Kortum, Ryungha Oh, Mark Rosenzweig, Nicholas Ryan, Matthew Schwartzman, Sam Slocum, Michael Sullivan, Kaushik Vasudevan, Trevor Williams as well as a host of seminar participants and visitors at Yale for their helpful comments and suggestions. 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 Yale Economic Growth Center that made this project possible.

†Yale University. lucas.conwell@yale.edu

I. Introduction

Workers in low- and middle-income countries waste significant amounts of time commuting (OECD 2016). One potential reason is that, in the face of limited fiscal capacity and increasing spatial fragmentation, governments often struggle to expand public transit at pace with urbanization (UN 2018; Angel et al. 2016). In its focus on public transit, the literature has so far neglected the privately-operated minibuses which provide enormous shares of urban transit in many cities: 50% in Cape Town, 89% in Lagos, and 98% in Dar es Salaam (Tun and Hidalgo, n.d.). These chaotic networks provide broad connectivity at the cost of long and unpredictable wait times – up to one-third of the typical commute in my context of Cape Town, South Africa – and poor personal safety. I study the extent to which this under-regulated *privatized shared transit* market reflects an efficient "invisible hand" at work or, as local officials surmise, gives rise to "an oversupply of minibus[es]" (Kerr 2018).

In particular, I pose two questions: which externalities exist in the privatized shared transit market, and how can policymakers intervene to improve the market allocation? Three features of shared transit likely render private provision inefficient. First, minibuses load passengers in a process that resembles a decentralized matching market. When individual minibuses enter the market, they internalize neither the lower passenger wait times to board nor the congestion in loading to which they contribute. Depending on which spillover prevails, the market equilibrium could yield a minibus quantity larger or smaller than optimal. Second, because minibuses share publicly-owned stations, the free-rider problem contributes to the under-provision of security services on-site. Third, minibuses suffer traffic congestion once en route.

To evaluate the efficiency of privately-provided transit, I build a micro-founded model of privatized shared transit subject to externalities in matching between buses and passengers, security provision, and road congestion. I estimate the model with newly-collected primary data on minibus operations in Cape Town and stated user preferences for exogenously-varied commute attributes. The quantified model reveals that Cape Town has too few, rather than too many, minibuses: an optimal subsidy on minibus entry increases welfare. In particular, considerable reductions in wait times for low-skill workers on long minibus routes outweigh slightly higher median wait times. However, the welfare gains from government-provided security guards at minibus stations dwarf those from either the aforementioned entry subsidy or the construction of exclusive minibus road lanes.

Two key features of my model formalize the notion of privatized shared transit. First, minibuses freely enter and match with passengers, and second, commuters with heteroge-

neous incomes optimally choose a mode of transport based on factors such as commute times and safety. At the heart of the model, a frictional matching market between minibuses and passengers determines the wait times of each. In particular, passenger wait times comprise two components: passengers first wait in long lines to board buses and subsequently wait on these buses, which depart only when full. Crucially, the number of buses affects these wait times in opposite ways. "Off-bus" wait times fall, and "on-bus" wait times rise, with minibus entry due to opposing thick-market and congestion externalities in matching. When the thick-market, or *boarding*, externality outweighs the congestion, or *filling*, externality, a sub-optimally low number of minibuses enters the market.

To quantify the model, I collected two forms of primary data in Cape Town. First, enumerators tracked passenger and bus queues on a stratified random sample of 44 minibus routes throughout a single morning commute. The raw data enables me to measure bus loading rates and commuters' wait times off and on the bus. Second, I introduce a new strategy, namely stated preference surveys, to generate exogenous variation in commute choices. In my survey, 526 respondents chose among hypothetical minibus commute options with exogenously-varied travel time, cost, and quality improvements, such as security. My survey, primarily of minibus commuters, identifies relative preferences for different minibus attributes. Preferences across modes, not only minibuses but also car and "formal" public transit, come from a separate stated preference survey conducted by the City of Cape Town.

With the help of these two datasets, I estimate the minibus matching function and the commuter demand system. In the queue data, the relationship between the relative number of waiting passengers and bus loading rates on a given route over time identifies the matching elasticities. The primary threat to identification comes from time-varying shocks correlated with routes' matching efficiencies, such as weather. Thus, I instrument for demand with pre-existing commute start times, measured in 2013. Since my queue data comes from 2022, the instrument satisfies the exclusion restriction as long as hourly matching efficiency trends do not strongly persist over a matter of years. From commuters' stated preferences for exogenously-varied attributes, I estimate a discrete choice model that yields commuters' mode-specific utility costs as well as values of time and quality improvements in dollar equivalents. Notably, high-skill commuters dislike and thus avoid minibuses due to security risks.

Finally, I employ the estimated model to analyze counterfactual policy strategies to remedy externalities in matching, security, and road congestion. First, there are too few minibuses in Cape Town. In fact, an optimally-set 36% entry subsidy gains low-skill commuters 0.73%

in welfare terms. Sizable gains accrue to commuters who travel between far-flung suburbs, where the boarding externality previously fostered under-provision of minibuses, which more than outweighs increases in wait time on the median route. Second, improvements in matching efficiency fail to provoke the same beneficial entry response on long routes. Third, low-skill commuters gain three times as much from government-provided security guards, even net of guard wages, as from the optimal subsidy. Fourth, exclusive lanes free minibuses from road congestion and produce similar net welfare gains as the entry subsidy. Thus, improved privatized transit could help solve the transportation problems of a host of rapidly growing, resource-poor African cities similar to Cape Town.

The policies for which I find significant welfare gains correspond directly to the recommendations of city planners and engineers. First, these experts discuss entry subsidies as a tool to flexibly serve newly-built housing developments on the "sparsely populated periphery" (Joubert 2013). In particular, predictions that subsidies will expand the range of profitable routes (Kerr 2018) directly anticipate my results. Second, Cape Town commuters report strong concerns over personal safety (Rayle 2017, 208–211). Not surprisingly, an earlier stated preference study involving minibus *drivers* finds a strong preference for station security (Plano, Behrens, and Zuidgeest 2020). Third, traffic engineering models predict substantial net gains from exclusive minibus lanes (De Beer and Venter 2021). This planning literature, however, lacks the quantitative models to consider how buses and commuters respond to such policies in equilibrium. For example, subsidies might provoke heterogeneous changes in minibus entry and, thus, wait time on routes of different lengths.

Related Literature

Existing work in economics carefully considers the general equilibrium demand response to urban transit investments. A series of recent papers exploit newly-opened government-provided rapid bus or subway lines as natural experiments. They then study how the associated travel time reductions translate into job matching (Tsivanidis 2019), worker informality (Zarate 2019), or gentrification (Balboni et al. 2020; Warnes 2020) in developing-country contexts. In particular, these papers emphasize rich general equilibrium spatial sorting effects. However, their exogenous-commute-cost frameworks do not directly apply to policy questions whose impact depends on privatized transit operators' decisions. In contrast, I contribute the first model of the privatized transit sector and formalize the responses of passengers' wait times, fares, and travel times to bus entry, demand, and road traffic.

Another strand of the literature endogenizes transport costs to compare the gains from

car- or public-transit-focused policies. These papers focus on road congestion (Allen and Arkolakis 2022; Barwick et al. 2022), network effects, and environmental externalities (Almagro et al. 2022). Similarly, in my framework, minibus and car travel times depend on the total traffic of both modes on each link traversed in the road network. My supply-side model, however, endogenizes a range of market outcomes beyond road congestion. For example, opposing externalities in minibus loading determine the passenger wait times that empirically account for almost one-third of minibus commute time, and prices respond to demand. In contrast to the existing literature, I can thus evaluate policies that specifically target privatized shared transit.

Finally, Brancaccio, Kalouptsidi, and Papageorgiou (2020) model ocean shipping with one-to-one matching and Poisson arrival shocks. I start with these two features and then develop a custom-designed model to fit the particulars of commute mode choice and minibus operations. On the latter front, for example, I replace one-to-one with many-to-one matching.

I structure the remainder of the paper as follows. In Section II, I describe my data, both newly collected and from existing sources, and the context. In Section III, I discuss a series of facts to rationalize my modeling choices and counterfactuals. I then lay out my theory in Section IV, followed by the estimation procedure in Section V. After validating my model's fit in Section VI, I discuss the welfare gains from alternative minibus 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, with additional details in Appendix B. I then provide an overview of the minibus market in Cape Town.

Newly-Collected and Household Survey Data

Little micro-data on privatized transit exists, so I collaborated with the mobility advisory firm GoAscendal to design a custom two-part data collection effort in Cape Town. First, minibus "station counts" tracked the matching-like process by which minibuses on a given route load at designated lanes in stations. For each route, enumerators recorded bus arrival and departure times, the number of passengers on board each bus, and the length of the queue to board at five-minute intervals. I later employ these records to calculate passenger and bus wait times along with the *loading rate* at which buses fill up. The counts covered a

FIGURE 1. STATED PREFERENCE SURVEY CHOICE SET BETWEEN MINIBUS OPTIONS

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

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

two-stage cluster sample of $N = 44$ minibus routes in Cape Town, where the eight clusters sampled in the first stage corresponded to the routes that originate from a given station. I stratified the station and the second-stage route samples by a proxy for the number of entering buses to reduce the number of zero bus observations.

Second, I designed a stated preference survey to estimate commuters' monetary values of quality improvements and time as well as cost sensitivity. I asked respondents to consider a hypothetical work commute trip and then choose a preferred option in a series of *choice sets* composed of two minibus alternatives, as in Figure 1. Each option varied exogenously in cost and travel time as well as in the presence or absence of three quality improvements, chosen based on commuter concerns in past city surveys. The latter included security guards at the publicly-owned, shared minibus stations, driver adherence to speed limits, and whether the minibus loads more passengers than seats. Respondents completed one of two randomized "blocks" of five choice sets; I chose the levels of the attributes with a *d-efficiency* algorithm that maximizes parameter precision (Rose and Bliemer 2009). After a pilot survey, I reduced the number of choice sets and alternatives per set to maintain respondent attention.

To conduct interviews, enumerators randomly approached respondents at one intermodal transport hub and two minibus stations on weekdays in June 2022. Resource constraints

precluded interviews at a greater variety of locations that would have avoided the resulting over-representation of minibus commuters in my sample (see Appendix B.2.3). Therefore, I employ this data only in the estimation of relative preferences for different *minibus* attributes and later attempt to quantify any disparities in preferences between my sample and the population.

Finally, I make use of representative household survey data collected by the City of Cape Town on stated preferences for different modes of transport. The city-run stated preference survey omitted quality improvements but varied the mode: car, various types of formal public transit, or minibus, as in Appendix Figure B.5. Additionally, for a larger sample of residents, the same survey provides home and workplace locations as well as incomes.

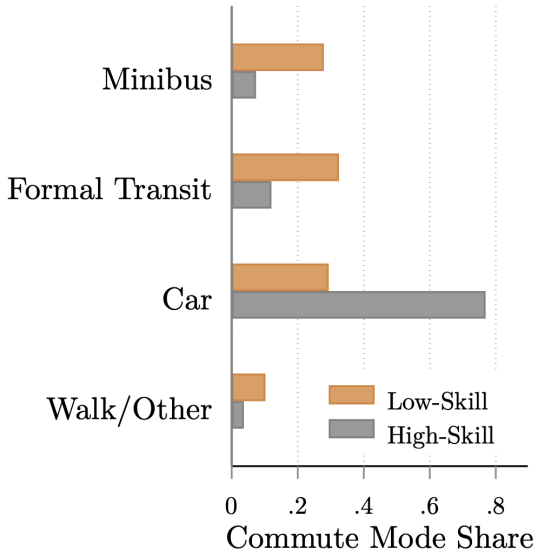
Minibuses in Cape Town

Privatized shared transit in South Africa takes the form of ubiquitous 15-passenger minibuses. Figure 2a displays the shares of commuters in Cape Town that use each mode of transport. Here and throughout the theory, I distinguish between two skill groups: *low*, or those with less than a college education, and *high*. A full 28% of low-skill workers and 7% of high-skill workers commute via minibus; the overwhelming majority of high-skill commuters instead drive to work. Around one-third of low-skill commuters use limited publicly-provided "formal" transit alternatives, including infrequent Golden Arrow buses that run in mixed traffic. However, the higher-speed network of MyCiti bus rapid transit and Metrorail train lines, overlayed on recent population growth in Figure 2b, misses many fast-growing suburbs.

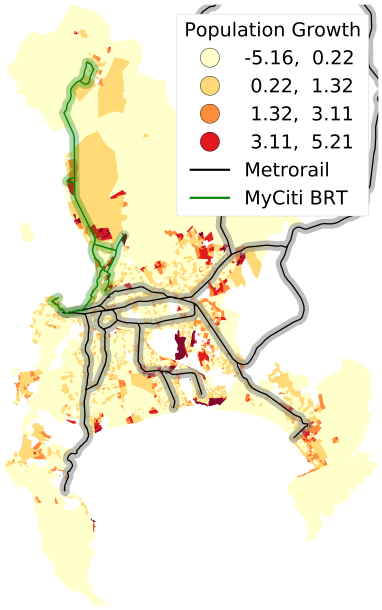
The minibus market comprises many small firms, each of whom, in one sample, owns an average of 1.87 buses (Woolf and Joubert 2013). Each firm pays an entry fee to an owner cooperative, or *association*, to operate on one distinct route defined by origin and destination (*Operating Licence Strategy 2013-2018* 2014; Kerr 2018). Additionally, firms must obtain a government permit and operate a vehicle with one of several approved seat capacities (Jobanputra 2018, 290). However, up to half of firms lack these permits (*Operating Licence Strategy 2013-2018* 2014, 77), and even "legal" minibuses receive virtually no direct government subsidy (Woolf and Joubert 2013). Minibus owners have little autonomy post-entry: associations set fares (Kerr 2018), subject to government approval of the "cost to the user (portion of monthly income spent on public transport)" (*Operating Licence Strategy 2013-2018* 2014, 65).

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-college) and high-skill commuters in Cape Town who use each mode, from 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 their commutes. 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.

III. FACTS ABOUT THE MINIBUS MARKET

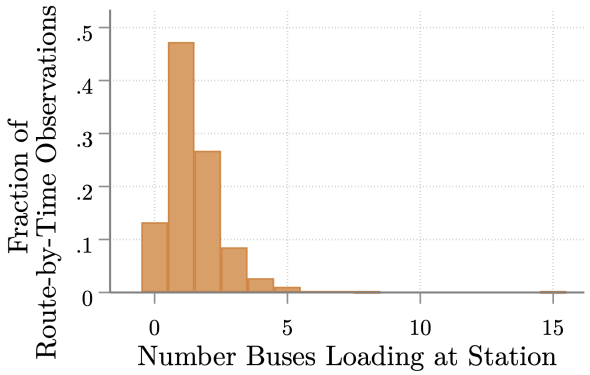
I now present six facts about the minibus sector in Cape Town. The first three facts each motivate a specific modeling choice. Routes begin at a so-called *taxi rank*: a large covered minibus station with clearly-marked loading lanes for each route. Henceforth, I deviate from local terminology and call these facilities *minibus stations* to avoid confusion.

Fact 1. *Multiple minibuses load simultaneously for each route, and large numbers of passengers wait to board, at minibus stations.*

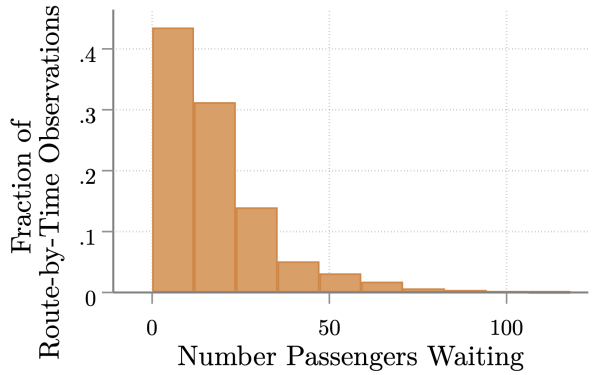
The histogram in Figure 3a displays the number of buses on the *same* route loading simultaneously in the station, across five-minute time blocks and routes in my station count data. Multiple buses load simultaneously in about half of the route-by-time observations, as pictured in Figure 3d. Furthermore, the histogram of the corresponding number of waiting passengers in Figure 3b demonstrates that passengers wait in queues that can exceed 50 people, echoing the visual in Figure 3c. In practice, passengers board in a random

FIGURE 3. MINIBUS ROUTE LOADING PROCESS AT ORIGIN STATION

(A) NUMBER OF BUSES LOADING



(B) NUMBER OF PASSENGERS WAITING



(C) LONG PASSENGER QUEUES TO BOARD

(D) MULTIPLE BUSES LOAD ON SAME ROUTE

(E) BUS CONGESTION AT STATION

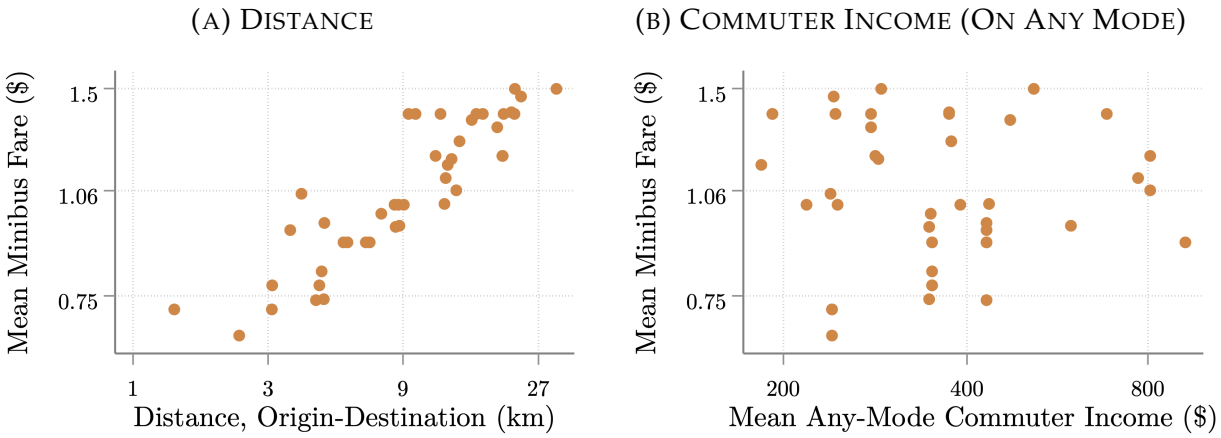
Notes: Panel (A) displays the distribution of the number of minibuses loading at the origin station on a specific minibus route, over minibus routes and five-minute periods. Panel (B) displays the distribution of the number of passengers waiting at the origin station for a specific route, also over routes and five-minute periods in my station count data from Cape Town. Panels (C)-(D) display images of the minibus loading process for a single route, defined by origin and destination and with a designated loading lane at the origin station, here the Cape Town CBD station. Panel (E) displays the exits from these loading lanes. Images by author, June 2022.

scramble. To replicate the large numbers of buses and passengers that wait simultaneously as well as the lack of an orderly queue, I model the minibus loading process as random matching between buses and passengers.

Fact 2. *15-passenger vans account for 94% of licensed minibuses, and 96% of minibuses depart with a full load of at least 15 passengers.*

Though the law allows for several discrete bus sizes, the 15-passenger-plus-driver variant dominates in practice (Jobanputra 2018, 290), so I impose a single exogenous bus size in my model. Because buses in my station count data virtually never leave less than full, as visualized in Appendix Figure A.2, I further require buses to reach this exogenous capacity

FIGURE 4. MINIBUS FARES VERSUS DISTANCE AND COMMUTER INCOME



Notes: Scatterplots display, all on log scales, the mean fare on a minibus route in my Cape Town sample, versus straight-line distance from the route's origin to its destination in Panel (A) and versus the mean income of commuters (using any mode) from the transport analysis zone (TAZ) where the route originates to the TAZ of its destination in Panel (B). Fares come from my on-board tracking data, and incomes come from the 2013 Cape Town Household Travel Survey.

before they depart.[1]

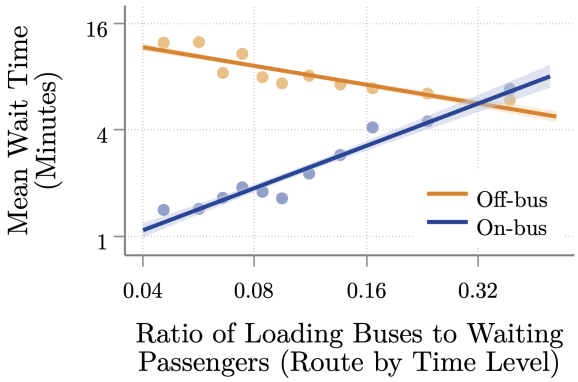
Fact 3. *Minibus fares increase strongly with route distance but not with commuters' ability to pay.*

Figure 4a plots distance on the horizontal axis and the mean fare charged on a minibus route in my route sample on the vertical axis; fares increase almost log-linearly with distance. Figure 4b instead displays the mean income of commuters, using any mode, between the route's origin and destination neighborhoods on the horizontal axis. Surprisingly, the fares on the vertical axis bear no apparent relationship to income. Fares thus appear to follow not from local demand conditions but instead from a citywide scheme. Motivated by Figure 4 and the aforementioned legal constraints on prices, minibus fares in my model adjust according to a reduced-form function of distance and citywide minibus demand.

The latter three facts rationalize my focus on specific externalities and counterfactual policies to correct them. I begin with the problem of passenger wait times. The long *off-bus* wait times implied by Fact 1 motivate local journalists to complain that "queues, especially during certain times of the day, are impossibl[y]" long (Theway 2018). Furthermore, the World Bank laments the "inefficien[cy]" of the behavior cited in Fact 2, namely "that minibus taxis generally only leave when they are full" (Kerr 2018). As a result, after

[1]The passenger experience once underway more closely resembles that of scheduled transit. Most routes follow a "line-haul" mode of operation during peak hours: they travel on highways to their destination and do not expect to pick up additional passengers.

FIGURE 5. OFF- AND ON-BUS PASSENGER WAIT TIMES VERSUS MINIBUS ENTRY



Notes: Binned scatterplots and best fit lines display the log-scale relationship between the relative number of loading buses to passengers waiting at the station for a given minibus route, across routes and five-min. periods in the station count data, and mean off-bus (orange) and on-bus (blue) passenger wait times.

boarding, passengers continue waiting *on-bus* for departure. In my Cape Town station count sample,

Fact 4. *The average minibus passenger waits 8.5 minutes off-bus to board and 2.7 minutes on-bus for departure, out of an average 36-minute minibus commute.*

I thus select counterfactuals that specifically target these wait times.

Fact 5. *Minibus passengers' off-bus wait time falls and on-bus wait time rises, the greater the number of loading minibuses relative to the number of waiting passengers on a route.*

On the horizontal axis, Figure 5 plots the number of loading buses relative to waiting passengers for a given route, computed every five minutes across routes in my station counts. The vertical axis displays minibus passengers' mean off- and on-bus wait times. Note first that higher numbers of loading buses correspond to lower off-bus wait times since passengers more easily find seats on buses. Bus entry, however, increases passengers' on-bus wait times between the moment they board and when the bus departs for its destination. This negative entry effect stems from passenger confusion, which slows the loading process, and in-station congestion among buses, as pictured in Figure 3e. Given these relationships, I consider a range of entry taxes and subsidies to reduce inefficiently long passenger waits.

Fact 6. *Commuters in a past satisfaction survey rated security second-most negatively among all minibus attributes.*

In a 2013 survey, security led the list of minibus-related grievances that included road

safety, crowdedness, cleanliness, timetable adherence, ease of use, and distance to the stop (see Appendix Figure A.3). Minibus associations could, of course, post security guards at the publicly-owned, shared minibus stations. However, because any benefits would spill over to routes that load nearby, the market will under-provide security. Commuters' complaints and this potential for under-provision thus motivate a counterfactual policy where the government finances minibus station security guards.

IV. A Theory of Privatized Shared Transit

In this section, I build a model of the privatized shared transit sector. My model has two key features: first, minibuses freely enter and match with passengers, and second, commuters with heterogeneous incomes optimally choose a mode of transport. I lay out the environment, discuss and solve the problems of each type of agent, and then define equilibrium.

Environment

I consider a city made up of a finite number of locations $i \in \{1, \ldots, I\} \equiv \mathscr{L}$ and two types of agents: minibuses and commuters. Time in my dynamic model is continuous, and commuters discount the future at rate r. Minibuses freely enter each origin-destination *route* ij and complete multiple trips, subject to matching frictions. Each location i spawns fixed masses $\left\{N_{ij}^{g}\right\}$ per unit time of commuters of skill $g \in \{\text{low } (l), \text{high } (h)\}$ who work in location j. Commuters choose one mode $m \in \{\text{minibus } (M), \text{formal transit } (F), \text{car } (A)\}$ for a *single* commute to work and collect a present-value wage ω_j^g upon arrival.

Minibus and car travel times depend on congestion in the road network, composed of all locations $\mathscr{L}$ as nodes and links between each location i and some fixed set of neighbors $L(i)$. I leave the remaining parameters related to formal transit and cars as exogenous. I solve the model exclusively in steady-state.

Minibuses

Minibuses enter every origin-destination pair, or what I term a route, ij at a cost $F_{ij} \equiv \overline{\psi} b_{ij}^{\phi}$. The cost "intercept" $\overline{\psi}$ accounts for entry fees paid to minibus associations as well as the monetary costs of bus purchase. However, the theory and calibration do not depend on the importance of each component. The elasticity ϕ summarizes the extent to which entry costs rise with the mass of loading buses, b_{ij}, on a route. For example, in my counterfactuals,

associations might charge higher entry fees on increasingly lucrative routes, as I micro-found in Appendix C.6, or introduce quantity restrictions.

After entry, minibuses complete multiple trips on the same route, as follows. First, minibuses load passengers at the minibus station in a frictional matching process. Importantly, I assume that buses depart only upon reaching an exogenous passenger capacity $\overline{\eta}$. Second, they collect fares τ_{ijM} from passengers. In line with Fact 3, fares adjust in counterfactuals according to a reduced-form fare function h of route straight-line distance, $\overline{\Delta}_{ij}$, and the aggregate minibus demand inflow $\overline{N}_M$. This approach flexibly captures market forces, as I micro-found with a model of minibus association fare choice in Appendix C.6, as well as political constraints. Third, the minibus travels through the road network toward its destination, subject to a per-kilometer operating cost χ.

In equilibrium, congestion between car commuters and minibuses on each road link traversed on the way to j determines the Poisson rate d_{ij} at which the minibus arrives. Upon arrival, the minibus exits the market with a probability $x \equiv g\left(T_{ij}\right)$ that increases with the expected total trip time T_{ij}, i.e., $g'\left(\cdot\right) > 0$. This structure incorporates the tradeoff between the number of trips and trip time inherent to a finite work shift and simultaneously preserves the stationarity of the model. Buses that do not exit costlessly return to the route origin i and begin another trip.

Matching

Each trip starts at a minibus station, where minibuses and passengers meet in a matching market segmented by route. The idea of incorporating matching frictions into a model of transportation comes from Brancaccio, Kalouptsidi, and Papageorgiou (2020), who consider ocean shipping. I extend their one-to-one to a many-to-one matching framework approximating the minibus loading process, where large crowds jostle to board one of several buses. Given a mass p_{ij} of waiting passengers and a mass b_{ij} of loading buses, $\mathscr{M}_{ij}$ matches between a minibus and a single passenger form per unit time, where

$$\mathscr{M}_{ij} \equiv \mu_{ij} p_{ij}^{\alpha} b_{ij}^{\beta}. \tag{1}$$

The exogenous matching efficiency μ_{ij} reflects station infrastructure and factors such as climate. In turn, the passenger and bus matching elasticities α and β capture the degree to which passengers rival each other and the extent of congestion among loading minibuses, respectively. The relative magnitudes of these two elasticities later determine how strongly minibus entry decreases passengers' off-bus wait times and increases their on-bus wait

times.

As in any matching model, each additional minibus that enters a route imposes two externalities on passengers and fellow buses. On the one hand, through the usual thick-market, or what I term *boarding*, externality, bus entry increases the boarding rates at which passengers meet buses,

$$\lambda_{ij} \equiv \frac{\mathscr{M}_{ij}}{p_{ij}} = \mu_{ij} b_{ij}^{\beta} p_{ij}^{\alpha-1}. \tag{2}$$

This positive externality is more significant, the less potent bus congestion relative to passenger congestion, i.e., the higher the bus matching elasticity β compared to the passenger elasticity α. On the other hand, through a negative congestion, or *filling*, externality, bus entry decreases the loading rate ι_{ij} at which buses meet passengers, given by

$$\iota_{ij} \equiv \frac{\mathscr{M}_{ij}}{b_{ij}} = \mu_{ij} p_{ij}^{\alpha} b_{ij}^{\beta-1}. \tag{3}$$

Lower levels of bus congestion go hand-in-hand with higher relative values of the bus matching elasticity β and shrink the magnitude of the filling externality. Thus, though the two externalities work in opposite directions, they do not necessarily cancel one other, even under constant returns to scale in matching.

As I now highlight, these two externalities affect passenger wait times in opposite ways. I rewrite the expected total passenger wait time on a given route as

$$\text{total wait}_{ij} \equiv \underbrace{\frac{1}{\lambda_{ij}}}_{\text{off-bus}} + \underbrace{\frac{1}{2}\frac{\overline{\eta}}{\iota_{ij}}}_{\text{on-bus}} = \underbrace{\left(\frac{b_{ij}}{p_{ij}}\right)^{-\beta}}_{\substack{\text{boarding} \\ \text{externality}}} + \underbrace{\frac{\overline{\eta}}{2}\left(\frac{b_{ij}}{p_{ij}}\right)^{1-\beta}}_{\substack{\text{filling} \\ \text{externality}}}, \tag{4}$$

where the second equality relies on $\mu_{ij} = 1$ and constant returns in matching. The first term in (4), or expected off-bus wait time, equals the inverse of the boarding rate. The second term, the expected on-bus wait time, equals half the time $\overline{\eta}/\iota_{ij}$ required for the bus to fill because passengers, in expectation, board half-full buses. Now, consider an increase in the relative number of loading buses, b_{ij}/p_{ij}. The higher this ratio, the lower passengers' off-bus wait time, precisely the boarding externality visible in the data in Figure 5. Simultaneously, the lower loading rates associated with additional entry increase the expected on-bus wait time. The former positive externality grows, and the latter negative filling externality falls in magnitude, the higher the bus matching elasticity β

relative to the passenger elasticity α. Thus, bus entry has an ambiguous effect on total passenger wait time. Potentially sizable heterogeneity in route profitability nevertheless precludes a direct mapping from elasticities to under- or over-provision, as I now describe.

Profits and Free Entry

In turn, lower bus loading rates reduce minibus profits by increasing total trip time. More precisely, the expected total time T_{ij} required to complete one trip on a route ij equals the sum of the time a bus waits to load and the expected travel time from i to j,

$$T_{ij} \equiv \frac{\overline{\eta}}{\iota_{ij}} + \frac{1}{d_{ij}}. \tag{5}$$

Loading time equals bus capacity $\overline{\eta}$ divided by the loading rate ι_{ij}, while average travel time equals the reciprocal of the Poisson destination arrival rate d_{ij}. The number of trips completed post-entry then follows a geometric distribution whose expectation, $\left[1 - g\left(\cdot\right)\right] / g\left(\cdot\right)$, decreases in the exit rate function $g\left(\cdot\right)$ and thus in average trip time, as in a model with a fixed work shift length.[2] Net revenue per trip, in turn, equals fare revenue $\overline{\eta}\tau_{ijM}$ minus the product of operating cost χ and the exogenous non-straight-line distance Δ_{ij} driven. Finally, expected minibus profits equal the product of the expected number of trips and net revenue per trip.[3] Free entry then ensures a bus loading rate, an expected trip time T_{ij}, and thus an expected number of trips, such that expected profit equals entry costs:

$$\underbrace{\frac{1 - g\left(T_{ij}\right)}{g\left(T_{ij}\right)}}_{\text{expected number trips}} \underbrace{\left(\overline{\eta}\tau_{ijM} - \chi\Delta_{ij}\right)}_{\text{net revenue per trip}} = \underbrace{\overline{\psi} b_{ij}^{\phi}}_{\text{entry cost}} . \tag{6}$$

Importantly, free entry generates considerable heterogeneity in passenger wait times across routes. In particular, a route's profitability decreases with length, measured in terms of either distance Δ_{ij} or the travel time component $1/d_{ij}$ of total trip time. Conditional on fares, minibuses on these longer routes must enjoy faster loading rates (ι_{ij}), requiring lower bus entry (b_{ij}). Any policy that alleviates the resulting high off-bus wait times on these long, ex-ante unprofitable routes might thus significantly increase commuter welfare.

[2]Note that I assume that buses do not receive the revenue from their last trip, so that the number of trips corresponds to a geometrically-distributed variable, conventionally used to model the number of failures until the first success. Thus, the expected number of revenue-generating trips satisfies $E\left[\#\text{ trips}\right] = \frac{1-x}{x} = \left[1 - g\left(T_{ij}\right)\right] / g\left(T_{ij}\right)$.

[3]Recall that buses decide to enter–and fares adjust according to h–but the former make no further choices thereafter.

Commuters

Commuters of skill g choose the mode m, either minibus, formal transit, or car, that offers the highest utility:

$$U_{ijm}^{g} - \kappa_{m}^{g} + \nu\varepsilon_{m}. \tag{7}$$

Total mode utility comprises (i) the deterministic commute value U_{ijm}^{g}; (ii) a mode-specific utility cost κ_{m}^{g}; and (iii) a Gumbel-distributed idiosyncratic preference ε_{m} with variance scaled by the parameter ν.[4]

The deterministic commute value U_{ijm}^{g} summarizes a multi-state process that varies by mode. Minibus commuters progress through three states: first, they wait off-bus to match at the minibus station; second, they wait on-bus for departure; and third, they enter a traveling state until their bus receives the Poisson arrival shock at rate d_{ij}. Formal transit commuters meet a vehicle at an exogenous rate λ_{ijF} and pay fares τ_{ijF}; unlike minibuses, formal transit immediately departs and arrives at its destination at an exogenous rate d_{ijF}. Car commuters pay a fixed monetary cost τ_{A} and reach their destination at the same congestion-affected Poisson rate as minibuses. In turn, the skill-specific utility cost κ_{m}^{g} represents commuters' non-pecuniary taste for a mode. Moreover, government actions to improve the quality of a mode – for example, the provision of security guards at minibus stations – could reduce utility costs.

I now characterize minibus passengers' value functions. Their deterministic commute value, or value of waiting off-bus, U_{ijM}^{g} follows an HJB equation:

$$rU_{ijM}^{g} = \lambda_{ij}\left[E_{n}\left(\tilde{U}_{ij}^{g}(n)\right) - U_{ijM}^{g}\right]. \tag{8}$$

Because passengers enjoy no flow value of waiting, the right-hand side equals the product of the rate λ_{ij} at which passengers board minibuses and the change in value from boarding a minibus.[5] The latter depends on the expectation of the value $\tilde{U}_{ij}^{g}(n)$ of waiting on-bus, which is itself a function of the passenger mass n already on board because fuller buses depart sooner. The (annuitized) on-bus waiting value equals the product of the rate ι_{ij} at which the bus fills and the change in value from a higher on-board passenger mass n, $r\tilde{U}_{ij}^{g}(n) = \iota_{ij}\tilde{U}_{ij}^{g\prime}(n)$. Furthermore, my timing assumptions require that the passenger

[4] I do not allow workers to combine modes; due to the rapidly declining quality of rail service in Cape Town, intermodal trips are rare. In my stated preference survey, for example, only 4.6% of minibus riders report also using formal buses or trains over the course of their commutes.

[5] I assume that commuters cannot observe the mass of passengers already on board, so they never reject a boarding opportunity.

value at the bus capacity $\overline{\eta}$ satisfies $\tilde{U}_{ij}^{g}(\overline{\eta}) = V_{ijM}^{g} - \tau_{ijM}$. Here, V_{ijM}^{g} denotes the traveling value, which increases in the destination arrival rate d_{ij} and the skill-specific wage ω_j^g received upon arrival according to $rV_{ijM}^{g} = d_{ij}\left(\omega_j^g - V_{ijM}^{g}\right)$. The formal transit and car value functions mirror those for minibus commuters and follow in Appendices C.1-C.2.

Aggregate commuter demand then adheres to three choice-probability equations of the familiar Gumbel form, which I derive through repeated value function substitution. First, skill-g commuters from home i to work j choose minibuses over formal transit or cars with a probability π_{ijM}^{g} that satisfies

$$\pi_{ijM}^{g} = \exp\left[\frac{\overline{W}_{ij}^{g}}{\nu}\right]^{-1} \exp\left\{-\kappa_M^g + \frac{\lambda_{ij}}{r+\lambda_{ij}}\left[1-\exp\left(\frac{-r\overline{\eta}}{\iota_{ij}}\right)\right]\frac{\iota_{ij}}{r\overline{\eta}}\left[\frac{d_{ij}\omega_j^g}{r+d_{ij}} - \tau_{ijM}\right]\right\}^{1/\nu}. \quad (9)$$

This minibus mode share thus depends on the associated utility cost κ_M^g, fare τ_{ijM}, and a series of "effective" discount factors applied to income, ω_j^g. These discount factors account for, in turn, the off-bus wait time through the boarding rate λ_{ij}, the on-bus wait through the loading rate ι_{ij}, and travel time through the arrival rate d_{ij}. The denominator is a function of commuters' ex-ante expected utility $\overline{W}_{ij}^{g}$.[6] Second, the formal transit choice probability, π_{ijF}^{g}, instead increases with the formal boarding rate λ_{ijF} as well as the destination arrival rate d_{ijF} and decreases with the utility cost κ_F^g as well as fare τ_{ijF}:

$$\pi_{ijF}^{g} = \exp\left[\frac{\overline{W}_{ij}^{g}}{\nu}\right]^{-1} \exp\left\{-\kappa_F^g + \frac{\lambda_{ijF}}{r+\lambda_{ijF}}\left[\frac{d_{ijF}\omega_j^g}{r+d_{ijF}} - \tau_{ijF}\right]\right\}^{1/\nu}. \quad (10)$$

Third, the car choice probability, π_{ijA}^{g}, depends on the utility as well as monetary costs, κ_A^g and τ_A, and arrival rate d_{ij}:

$$\pi_{ijA}^{g} = \exp\left[\frac{\overline{W}_{ij}^{g}}{\nu}\right]^{-1} \exp\left(-\kappa_A^g + \frac{d_{ij}}{r+d_{ij}}\omega_j^g - \tau_A\right)^{1/\nu}. \quad (11)$$

Finally, I illustrate how commuters value different commute attributes with a first-order approximation, around $r = 0$, to the minibus choice probability. Improved station security and other quality improvements would affect commuters' likelihood of choosing minibuses in (12) through κ_M^g. In turn, the *product* of the time preference rate r and the wage ω_j^g determines commuters' value of reductions in the total commute time in parentheses. The latter equals the sum of expected off- and on-bus wait as well as travel time. Finally, the

[6] The usual properties of the Gumbel distribution imply that $\overline{W}_{ij}^{g} \equiv \nu \log\left[\sum_m \exp\left[U_{ijm}^{g} - \kappa_m^g\right]^{1/\nu}\right]$.

inverse of the Gumbel shape parameter ν yields commuters' sensitivity to fares. In my stated preference survey, I exogenously vary quality improvements to κ_M^g as well as travel times to estimate commuters' monetary valuations thereof.

$$\pi_{ijM}^g \approx \exp\left[\frac{\overline{W}_{ij}^g}{\nu}\right]^{-1} \exp\left[\underbrace{-\kappa_M^g}_{\text{utility cost}} - r\omega_j^g\left(\underbrace{\frac{1}{\lambda_{ij}}}_{\substack{\text{off-bus}\\\text{wait}}} + \underbrace{\frac{1}{2}\frac{\overline{\eta}}{\iota_{ij}}}_{\substack{\text{on-bus}\\\text{wait}}} + \underbrace{\frac{1}{d_{ij}}}_{\substack{\text{travel}\\\text{time}}}\right) - \underbrace{\tau_{ijM}}_{\text{fare}}\right]^{1/\nu}.^{7} \tag{12}$$

Road Congestion

Minibus entry might congest roads in addition to the loading process, a channel incorporated into minibuses' and cars' destination arrival rates d_{ij} as follows. The travel time over an individual link in the road network, $t_{ik} \equiv \bar{t}_{ik} v_{ik}^{\gamma}$, depends on the free-flow travel time $\bar{t}_{ik}$ and increases with the vehicle inflow v_{ik}, raised to a road congestion elasticity γ. Minibuses thus impose a lower per-commuter externality, given their higher capacity, than cars. Unlike Allen and Arkolakis (2022), I abstract away from route choice through the road network. For tractability, I instead require that minibuses and cars follow the sequence of links, or path, $\rho(i,j)$ that minimizes free-flow travel time from i to j, as in Appendix C.4. The destination arrival rate then equals the inverse sum of travel times over road network links along that path, $d_{ij} \equiv \left(\sum_{kk' \in \rho(i,j)} t_{kk'}\right)^{-1}$. Expected travel times, therefore, hinge on traffic not in the aggregate as in Barwick et al. (2022) but instead on *each link* traversed in the road network on the way to commuters' destinations.

Equilibrium

First, I summarize the equilibrium conditions that determine demand and supply. The former adheres to the commuter choice-probability equations (9)-(11). Two equations jointly characterize supply: the free entry condition (6) and the reduced-form fare function. As for the latter, aggregate minibus demand inflows satisfy $\overline{N}_M \equiv \sum_{i,j,g} N_{ij}^g \pi_{ijM}^g$, such that fares satisfy

$$\tau_{ijM} = h\left(\overline{\Delta}_{ij}, \sum_{i,j,g} N_{ij}^g \pi_{ijM}^g\right). \tag{13}$$

[7] Note that I have suppressed the term $r\tau_{ijM}\left(\frac{1}{\lambda_{ij}} + \frac{1}{2}\frac{\overline{\eta}}{\iota_{ij}}\right)$, which will be small due to the multiplication of a small interest rate and a small fare.

Second, I characterize the matching rates on each minibus route. In steady-state, the inflow to the stock, p_{ij}^g, of minibus passengers of skill g waiting off-bus equals the outflow, or $N_{ij}^g \pi_{ijM}^g = \lambda_{ij} p_{ij}^g$. The boarding rate, however, depends on the sum of (off-bus) waiting passengers across skill groups, $p_{ij} \equiv \sum_g p_{ij}^g$. These insights, combined with (2)-(3), yield equations for the passenger boarding rate,

$$\lambda_{ij} = \mu_{ij}^{1/\alpha} \iota_{ij}^{-\beta/\alpha} \left[\sum_g N_{ij}^g \pi_{ijM}^g \right]^{(\alpha+\beta-1)/\alpha}, \tag{14}$$

and bus loading rate,

$$\iota_{ij} = \frac{\sum_g N_{ij}^g \pi_{ijM}^g}{b_{ij}}. \tag{15}$$

Third, I employ the definitions of the road-based destination arrival rate d_{ij} and link-level driving time t_{ik} to establish, in Appendix C.4, that

$$d_{ij} = \left[\sum_{kk' \in \rho(i,j)} \bar{t}_{kk'} \left(\sum_{i'} \sum_{j'} \left[\mathbb{1}\left\{ kk' \in \rho(i',j') \right\} \sum_g \left(\frac{N_{i'j'}^g \pi_{i'j'M}^g}{\overline{\eta}} + N_{i'j'}^g \pi_{i'j'A}^g \right) \right] \right)^{\gamma} \right]^{-1}. \tag{16}$$

Equilibrium. *Given parameters* $\left\{ r, \nu, \boldsymbol{\kappa}, \alpha, \beta, \mu, \gamma, \tau_A, \overline{\psi}, \phi, \overline{\eta}, g(\cdot), \chi, h(\cdot) \right\}$ *and model geography* $\left\{ \boldsymbol{N}, \boldsymbol{\omega}, \bar{\boldsymbol{t}}, \boldsymbol{\lambda_F}, \boldsymbol{d_F}, \boldsymbol{\tau_F}, \boldsymbol{\Delta}, \overline{\boldsymbol{\Delta}} \right\}$*, an equilibrium is a vector* $\left\{ \boldsymbol{b}, \boldsymbol{\pi}, \boldsymbol{\tau_M}, \boldsymbol{\lambda}, \boldsymbol{\iota}, \boldsymbol{d} \right\}$ *such that (i) free entry of minibuses* (6) *holds; (ii) commuter demand is consistent with* (9)-(11)*; (iii) the function* (13) *determines fares; (iv) passenger boarding rates in* (14) *as well as (v) bus loading rates in* (15) *are consistent with matching technology; and (vi) road-based destination arrival rates reflect congestion* (16)*.*

V. Estimation of Minibus and Demand Parameters

I combine my newly-collected data with existing micro-data to estimate the model for a geography composed of the $I = 18$ *transport analysis zones* in Cape Town, mapped in Appendix Figure B.4.[8] The road network consists of each zone as a node along with links from each zone to those it borders.

[8] I do not include commuters living and working in the same transport analysis zone in my calibrated commute flows.

TABLE 1. CALIBRATED PARAMETERS

Parameter	Description	Value
Externally Calibrated		
I	Number Locations	18
N_{ij}^g	Commute Flows	
ω_j^g	Wages	
$\bar{t}_{ik}$	Free-Flow Driving Time	
λ_{ijF}	Formal Arrival Rate	
d_{ijF}	Formal Destination Arrival Rate	
τ_{ijF}	Formal Fare	
τ_A	Car Commute Cost	5.2
δ_0	Minibus Shift Length	240
δ_1	Minibus Inverse # Trips	0.01
χ	Per-km. Operating Cost	0.06
Δ_{ij}	Route Driving Distance	
$\overline{\Delta}_{ij}$	Straight-Line Distance	
Minibus Supply		
ϕ	Entry Cost Elasticity	0.0143
Γ_0	Fare Intercept	1.73
Γ_1	Fare Distance Slope	0.29
Γ_2	Fare Demand Slope	0.0438

Parameter	Description	Value
Minibus Matching		
α	Passenger Elasticity	0.65
β	Bus Elasticity	0.44
Stated Preference		
r	Commuter Rate of Time Pref.	0.001
ν	Gumbel Shape	4.76
κ_M^l	Low-Skill Minibus Util. Cost	7.7
κ_M^h	High-Skill Minibus Util. Cost	15
κ_F^l	Low-Skill Formal Util. Cost	3.6
κ_F^h	High-Skill Formal Util. Cost	9.2
Internally Calibrated		
$\overline{\psi}$	Minibus Entry Cost Intercept	33.5
$\overline{\eta}$	Minibus Capacity	6.7
μ	Minibus Matching Efficiency	0.24
Road Congestion		
γ	Road Congestion Elasticity	0.0917

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 Google Maps and Azure APIs (see Appendix D.5). The entry cost elasticity uses the station count data, and minibus fares are estimated using on-board tracking and National Household Travel Survey data. The minibus matching elasticities are estimated using the station counts. The stated preference estimation uses my new survey and an existing module from the aforementioned 2013 survey. The internal calibration minimizes the distance to moments in the station counts. Finally, the road congestion elasticity uses TomTom API data.

Structural Parameters

I estimate the model's structural parameters in five main steps. First, I devise an instrumental variables strategy in the station count data to identify the matching elasticities α and β. Second, from commuters' stated preferences, I estimate their mode-specific utility costs κ_m^g, rate of time preference r, and Gumbel shock shape ν. Third, I externally calibrate the geography and other secondary parameters. Fourth, I estimate the minibus entry cost elasticity ϕ and the reduced-form minibus fare function. Finally, conditional on all other parameters, I internally calibrate the minibus capacity $\overline{\eta}$, entry cost intercept $\overline{\psi}$, and matching efficiency μ. Table 1 summarizes all calibrated parameters.

Minibus Matching

First, I estimate the passenger and (mini)bus matching elasticities, α and β, using variation in demand over time within a given minibus route in my station count data. Across 44 routes, indexed by origin i and destination j, and 48 five-minute clock time intervals indexed by t, I estimate the empirical equivalent of equation (3) for the bus loading rate:

$$\log \iota_{ijt} = \alpha \log p_{ijt} + (\beta - 1) \log b_{ijt} + \overline{\mu}_{ij} + \overline{\mu}_t + \overline{\mu}_{it} + \epsilon_{ijt}.^{9} \tag{17}$$

I calculate the bus loading rate ι_{ijt} as the number of boarding passengers on a route per bus and minute; recall that p_{ijt} and b_{ijt} denote waiting passengers and buses, respectively.[10] Furthermore, I model matching efficiency, $\log \mu_{ijt}$, as a combination of route ($\overline{\mu}_{ij}$), time ($\overline{\mu}_t$), and origin-by-time ($\overline{\mu}_{it}$) fixed effects, as well as an idiosyncratic shock ϵ_{ijt} to account for out-of-steady-state dynamics in the data. In particular, the origin-time fixed effects control for any shocks that equally affect the matching efficiencies of all routes that originate from a given station. The vast majority of plausible threats to identification, such as localized rain or special events which might both slow the loading process and affect demand, likely fall into this category. Rain shocks would nevertheless bias OLS estimates of the loading rate equation (17) to the extent that they affect routes at the same station differently, perhaps because only some have roofs over their loading lanes.

To address this more remote possibility, I pursue an instrumental variables strategy. I assume constant returns, $\alpha + \beta = 1$, to rewrite the estimating equation as a function of the relative number of waiting passengers to buses, i.e., $\log \iota_{ijt} = \alpha \log \left(p_{ijt} / b_{ijt} \right) + \overline{\mu}_{ij} + \overline{\mu}_t + \epsilon_{ijt}$.[11] Then, I instrument for $\log \left(p_{ijt} / b_{ijt} \right)$ with the log number of commuters resident in i who reported, in a 2013 survey, that they leave around time t to go to work. Since I measured loading rates ι_{ijt} in the year 2022, minute-by-minute matching efficiency trends in ϵ_{ijt} would have to persist over nine years to violate the exclusion restriction. Even under such persistence, the distribution of commute start times remains exogenous provided commuters do not choose their residential locations or the start times of their working

[9] The theory suggests many equivalent matching function estimation methods. I choose equation (3) for ι_{ij} to permit estimation with a single instrumental variable as well as later validation against actual bus wait times.

[10] The 44 routes in my data originate in 6 of the 18 transport analysis zone (TAZ) units that correspond to model locations, and their destinations cover 14 TAZ. Together, these origins and destinations form 28 unique TAZ pairs. Of 1,917 observations with non-missing data for all three variables in the estimating equation, I lose 41 with a zero bus loading rate, 167 with nonzero bus loading rates but zero waiting passengers, and another 82 with nonzero loading rates and passengers but zero loading buses. Appendix B.1.3 provides details on the data construction.

[11] Because I have only one instrument, I cannot estimate flexible returns to scale in the IV specifications.

TABLE 2. MATCHING FUNCTION ESTIMATES

	OLS			*IV* with $\alpha + \beta = 1$	
	(1)	(2)	(3)	(4)	(5)
Parameter	log bus loading rate	log bus loading rate	log bus loading rate	log bus loading rate	log bus loading rate
α	0.687***	0.662***	0.645***	0.841***	0.665
Passenger Matching Elasticity	(0.0130)	(0.0188)	(0.0264)	(0.106)	(1.060)
β	0.433***	0.425***	0.435***		
Bus Matching Elasticity	(0.035)	(0.042)	(0.043)		
95% CI for $\alpha + \beta$	[1.03,1.21]	[0.98,1.19]	[0.97,1.19]		
Route FE	✓	✓	✓	✓	✓
Time FE		✓	✓		✓
Origin-Time FE			✓		
Observations	1,627	1,627	1,607	1,316	1,316
R-Squared	0.587	0.818	0.839	0.50	0.56
First-Stage F Statistic				14.80	0.24

Notes: Robust standard errors in parentheses, clustered at the origin level; *** indicates p<0.01, ** p<0.05, and * p<0.1. Columns 1–3 present estimates of (17) over five-minute time blocks and 44 routes in my station count data, with fixed effects included, as noted, for route, time, or origin station by time. In Columns 4–5, I assume constant returns to scale, $\beta = 1 - \alpha$, so that I can regress the log bus loading rate on the log ratio of the stock of waiting passengers to the stock of loading buses. I instrument for this ratio using the log number of commuters living in the mesozone spatial unit where the route originates who report leaving their home during the 15-minute period including time t, calculated from the 2013 Cape Town Household Travel Survey.

days based on within-route matching efficiency trends.

Table 2 displays the matching elasticities that result from the estimation of (17), first by OLS. In the specification in Column 1, I include only route fixed effects; in Columns 2 and 3, I add time and origin-time fixed effects, respectively. In my preferred specification in Column 3, I obtain a precisely estimated passenger elasticity $\hat{\alpha} = 0.645$, which is 50% larger than the bus elasticity, $\hat{\beta} = 0.435$. Jointly, these two estimates imply approximately constant returns to scale, so I can safely assume the latter in the instrumental variables specifications in Columns 4–5. The passenger elasticity $\hat{\alpha} = 0.841$, thus estimated in Column 4, is larger than in prior specifications, but the corresponding 95% confidence interval includes the OLS estimate. In Column 5, I include time fixed effects and obtain a noisy but broadly similar estimate. Furthermore, the estimates change little when, in Appendix D.1, I weight by routes' inverse sampling probabilities or estimate an alternative specification for the total, rather than per-bus, number of loading passengers. In my model calibration, I employ the OLS estimates from Column 3 due to their greater precision and to allow flexible returns to scale.

Stated Preferences

Second, stated preference data informs four sets of demand parameters, namely the four mode-skill-specific utility costs κ_m^g, the effects by skill group θ_z^g of three minibus quality improvements on utility costs, the rate of time preference r, and the Gumbel shock shape parameter ν. Stated preference surveys have recently proliferated in economics, for example, in the study of long-term care (Ameriks et al. 2020) or marriage decisions (Andrew and Adams-Prassl 2021). In the present context, I employ stated rather than revealed preferences because they provide exogenous variation in commute attributes (Ben-Akiva, McFadden, and Train 2019) as well as new measures of quality improvements (Carson and Czajkowski 2014). Furthermore, respondents make the short-term mode choices that correspond to my scenarios almost daily, so they could likely predict their hypothetical behavior accurately.

I first derive the model-implied utility of each stated preference alternative. In so doing, I begin with the utility cost, which I now allow to depend not only on the mode but also on the bundle of (binary) quality improvements offered by an alternative. Let $\overline{\kappa}_{cl}^g$ denote the utility cost of alternative l in choice set c for skill group g, and $m(c,l)$, the associated mode of transport. Because the stated preference data does not specify quality improvements for formal transit and car alternatives, their utility costs do not differ across alternatives, i.e., $\overline{\kappa}_{cl}^g = \kappa_{m(c,l)}^g$ for $m(c,l) \in \{F, A\}$. In the case of minibuses, each alternative specifies some combination of quality improvements, indexed by z, from among security, speed limit adherence, and lack of overcrowding. Let $q_{cl}(z)$ denote an indicator for the presence of quality improvement z in alternative l of choice set c. I then assume that the minibus utility cost satisfies $\overline{\kappa}_{cl}^g \equiv \kappa_M^g + \sum_z \theta_z^g q_{cl}(z)$. Here, κ_M^g denotes the no-quality-improvement minibus utility cost, and θ_z^g is the effect of binary quality improvement z on the utility cost of skill group g. The linearly-approximated utility, net of individual fixed effects and the preference shock, of individual i of skill g from alternative l in choice set c equals

$$\overline{U}_{icl}^g \approx -\left(\kappa_{m(c,l)}^g + \sum_z \theta_z^g q_{cl}(z)\right) - r\omega_i\left(w_{cl} + t_{cl}\right) - \tau_{cl} + rw_{cl}\tau_{cl}, \tag{18}$$

This utility depends not only on utility costs but also on individual income ω_i, the indicated wait as well as travel times, w_{cl} and t_{cl}, and fares τ_{cl}.

Since I assume Gumbel-distributed idiosyncratic preference shocks, commuters' mode choices, based on (18), directly correspond to a multinomial logit discrete choice model. I estimate this model on a stacked dataset, composed of stated choices from my own and the

city-run survey, using maximum likelihood, with alternative-l choice probabilities given by

$$\pi^g_{icl} = \frac{\exp\left[\zeta^g_{m(c,l)} + \sum_z \beta^g_z q_{cl}(z) + \beta_{\text{time}}\omega_i\left(w_{cl} + t_{cl}\right) + \beta_{\text{fare}}\tau_{cl} + \beta_{\text{resid}} w_{cl}\tau_{cl}\right]}{\sum_{l'} \exp\left(\overline{U}^g_{icl'}/\nu\right)}. \tag{19}$$

The ratio of the group-mode fixed effect $\zeta^g_{m(c,l)}$ to the fare effect β_{fare} places utility costs in dollar terms, where I normalize the car utility cost $\kappa^g_A = 0$. Crucially, while I over-sampled minibus commuters, my survey included only minibus alternatives and so does not contribute to the identification of the relative utility costs across modes. The choices of respondents in my data do, however, yield the effect of each quality improvement z on utility cost, calculated as $\theta^g_z = \beta^g_z/\beta_{\text{fare}}$. I obtain the rate of time preference r from the ratio $\beta_{\text{time}}/\beta_{\text{fare}}$, or the relative increase in the disutility of commute time $(w_{cl} + t_{cl})$ with income ω_i. Finally, commuters' inverse responsiveness, $|1/\beta_{\text{fare}}|$, to fares τ_{cl} identifies the Gumbel shape ν.

Table 3 displays the parameters estimated from the multinomial logit model in (19). I begin with the rate of time preference, $\hat{r} = 0.001$, which implies that commuters would sacrifice only 1% of their daily wage to save ten minutes of commute time to and from work. Next, I estimate a Gumbel shape parameter $\hat{\nu} = 4.76$ and, by implication, a low cost sensitivity: minibus relative choice probabilities would fall by only 24% in response to a 100% increase in fares. The low- and high-skill minibus utility costs, $\hat{\kappa}^l_M = 7.68$ and $\hat{\kappa}^h_M = 15.03$, demonstrate that both groups dislike minibuses relative to cars. If commuters experienced identical utility costs for minibuses and cars, their relative minibus choice probabilities would rise by a factor of 4.35 for low-skill workers and 23.5 for high-skill workers. The same is true for formal transit, albeit to a lesser degree.

In the right panel of Table 3, all three quality improvements significantly decrease minibuses' utility cost for the low-skill group; high-skill commuters place a premium on security but not (significantly) on speed limit adherence. For example, the provision of station security would almost double the high-skill minibus mode share relative to any other mode. To quantify any heterogeneity in preferences for minibus attributes, I re-estimate the logit model (19) among only respondents interviewed outside of minibus stations. Separately, I weight my survey by realized commute mode choices; the results, in Appendix D.2.2, change little.

TABLE 3. STATED PREFERENCE SURVEY ESTIMATES

Parameter	Value	
r	.001***	
Commuter Rate of Time Pref.	(.0004)	
ν	4.76***	
Gumbel Shape	(1.26)	
Utility Costs	*Low-Skill*	*High-Skill*
κ_F	3.63***	9.17***
Formal Transit Utility Cost	(0.51)	(1.89)
κ_M	7.68***	15.03***
Minibus Utility Cost	(1.56)	(3.55)

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

Notes: Robust standard errors in parentheses; *** indicates p<0.01, ** p<0.05, and * p<0.1. Estimates reflect $N = 19,712$ individuals by choice sets by alternatives in either my newly collected minibus stated preference survey or a stated preference module of the 2013 Cape Town Household Travel Survey. The estimated parameters come from a multinomial logit model with choice probabilities given by (19). I normalize $\kappa_A^g = 0$ and restrict the sample to individuals employed outside the home between 25 and 65 years of age.

External Calibration

Third, I externally calibrate secondary parameters as well as the model geography. Most importantly, I quantify the road congestion elasticity γ with data from TomTom's Traffic Stats API on traffic volume v_{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 $\log t_{ih} = \bar{t}_i + \gamma \log v_{ih} + \epsilon_{ih}$,; the unobserved shocks ϵ_{ih} that threaten identification include weather and special events (Barwick et al. 2022). In Column 1 of Appendix Table D.7, I obtain $\hat{\gamma} = 0.0917$. Next, I parameterize the function g that determines the rate x at which a minibus exits upon arrival at its destination as

$$x \equiv g\left(T_{ij}\right) \equiv 1/\left\{1 + \exp\left[-\delta_1\left(T_{ij} - \delta_0\right)\right]\right\}. \tag{20}$$

The workday "length" parameter δ_0 corresponds to the total time available to the bus to make trips. The parameter δ_1, then, determines how quickly the expected total trip time T_{ij} decreases the feasible number of daily trips. I calibrate these two parameters to produce reasonable expected numbers of trips for various trip times. Finally, I directly obtain the car commute cost τ_A, minibus operating cost χ, and the model geography from the data. The latter includes commute flows N_{ij}^g, wages ω_j^g, link-level free-flow driving times $\bar{t}_{ik}$ as well as origin-destination free-flow driving distances Δ_{ij}, and the formal transit system, $\left\{\lambda_{ijF}, d_{ijF}, \tau_{ijF}\right\}$. I provide additional details regarding all externally calibrated parameters

in Appendix D.5.

Minibus Entry and Supply

Fourth, I quantify the congestion elasticity ϕ of minibus entry costs with the same station count data used in the matching estimation. From free entry, I derive that bus loading times $\overline{\eta}/\iota_{ijt}$ on route ij at time t decrease in the number of loading buses b_{ijt}:

$$\frac{\overline{\eta}}{\iota_{ijt}} = \zeta_0 - \frac{\phi}{\delta_1} \log b_{ijt} + \zeta_{ij} + \zeta_{it} + \varepsilon_{ijt} \tag{21}$$

Route fixed effects, ζ_{ij}, and origin-time fixed effects, ζ_{it}, control for variation in fares and operating costs, so the threats to identification in ε_{ijt} include route-specific changes in travel time. I estimate the free entry equation (21) in Appendix Table D.4 and, conditional on the parameter δ_1 that controls the number of trips minibuses can accomplish per day, find an entry cost elasticity of $\hat{\phi} = 0.0143$. Congestion in entry, or quantity restrictions on the part of minibus associations, thus appear relatively minimal.

Fifth, I parameterize the reduced-form minibus fare function h as

$$\log \tau_{ijM} \equiv \log \left[h \left(\overline{\Delta}_{ij}, \overline{N}_M \right) \right] \equiv \Gamma_0 + \Gamma_1 \log \overline{\Delta}_{ij} + \Gamma_2 \log \left(\overline{N}_M \right). \tag{22}$$

The exogenous parameters $\{\Gamma_n\}$ quantify the intercept as well as the effects of straight-line route distance $\overline{\Delta}_{ij}$ and citywide minibus demand $\overline{N}_M$ on fares. Minibus tracking data on fares and distances informs Γ_1; I estimate Γ_2 with nationwide data on local minibus fares as well as demand and instrument for the latter with population. Appendix D.4 details the entry and fare estimation procedures.

Internal Calibration

Finally, I match three aggregate moments in the station count data: (i) the median, over routes and time periods, of the relative number of loading buses, b_{ijt}/p_{ijt}; (ii) the median bus loading time, $\overline{\eta}/\iota_{ijt}$; and (iii) the median off-bus passenger wait time, $1/\lambda_{ijt}$. 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 relative number of loading buses identifies entry costs $\overline{\psi}$, bus loading time identifies bus capacity $\overline{\eta}$, and the passenger off-bus wait time determines matching efficiency μ. The final column lists the calibrated parameter values. To interpret the results, I note that the median realized entry cost, which depends not only on the intercept $\overline{\psi}$ estimated in Table 4 but also on the number of buses,

TABLE 4. INTERNAL CALIBRATION

Moment			Parameter		
Description	*Data*	*Model*	*Description*		*Value*
Median Loading Buses/Waiting Passengers	0.09	0.09	$\overline{\psi}$	Entry Cost Intercept	33.5
Median Bus Loading Time	4	4	$\overline{\eta}$	Minibus Capacity	6.7
Median Off-Bus Passenger Wait Time	7.18	7.18	μ	Matching Efficiency	0.24

Notes: This table displays the moments used in internal calibration: medians across routes and five-minute periods in the station count data and medians across routes in the model. 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 and numerically minimize the sum of squared (percentage) deviations from the three moments.

exceeds the fare revenue from the median trip by a factor of four.

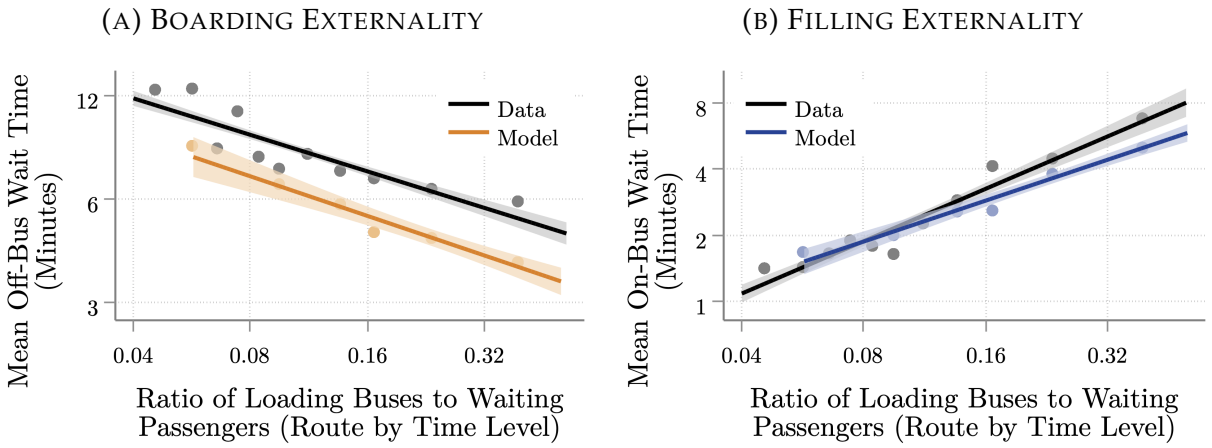
VI. MODEL-PREDICTED MINIBUS OPERATIONS AND AGGREGATE PATTERNS

I now demonstrate that my model matches non-targeted data: the minibus network, the boarding and filling externalities, and aggregate mode shares. First, the model-predicted bus entry on each origin-destination pair mirrors a rough proxy of entry in the data, namely the number of distinct minibus routes that link each pair of neighborhoods (see Appendix Figure A.4).[12] Importantly, the model matches the concentration of minibuses in central neighborhoods and their ability to directly link outlying suburbs, in contrast to the usual service patterns of publicly-provided transit. My framework, where commute flows as well as route length affect minibus entry, thus approximates the actual minibus supply successfully.

Second, the model replicates the boarding and filling externalities which play a central role in my theory. Figures 6a and 6b display the relative number of loading buses to passengers on the horizontal axis and off-bus or on-bus wait times on the vertical axis. I plot route-by-time observations in the station count data in black and routes in the model in orange. In both figures, my calibration targets the *medians* of the horizontal and vertical axes but not the relationship between bus entry and wait times. Reassuringly, the model nevertheless generates off-bus wait times that fall with bus entry in Figure 6a, precisely as in the data. I match the filling-externality relationship in Figure 6b even more closely.

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

FIGURE 6. MATCHING EXTERNALITY RELATIONSHIPS IN DATA VERSUS MODEL



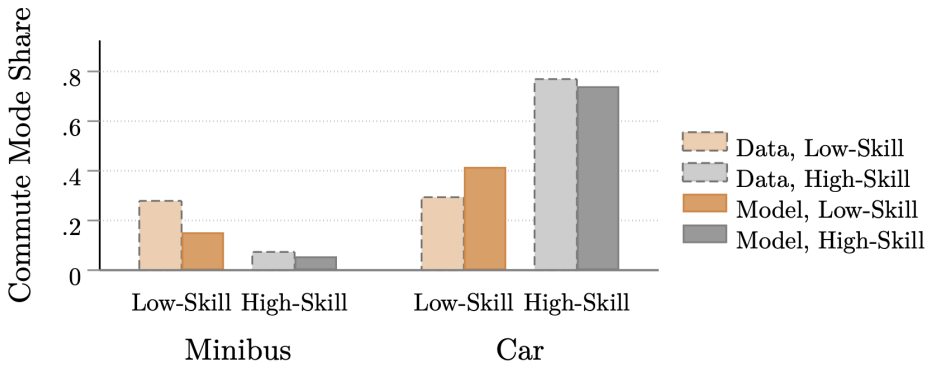
Notes: Panel (A) displays a binned scatterplot and best-fit line corresponding to the boarding externality of entry, i.e. the log-scale relationship between the relative number of loading buses to waiting passengers, $\log\left(b_{ij}/p_{ij}\right)$, across routes and five-min. periods in the station count data and across routes as predicted by the model, and expected passenger off-bus wait time, $\log\left(1/\lambda_{ij}\right)$. Panel (B) displays the filling externality, instead plotting expected passenger on-bus wait time, $\log\left[\overline{\eta}/\left(2\iota_{ij}\right)\right]$, on the vertical axis.

Third, the model accurately predicts aggregate commute choices. The bars in Figure 7 indicate the skill-group-level shares of Cape Town commuters in the household survey data (light colors) and model (dark colors) who use each mode. In both data and model, a sizable share of low-skill commuters choose minibuses, while high-skill commuters choose minibuses at comparatively low rates and cars at high rates. The close match between these non-targeted mode shares computed from 2013 data and those implied by the model suggests that the stated preference methodology accurately captures 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, laid out in Appendix E.1.4. First, the model predicts not only citywide mode shares, as in Figure 7, but also the reported commute modes of the stated preference respondents. Second, demographic heterogeneity in commuters' values of time and quality improvements largely follows intuition. Women, for example, place a higher value on time saved, as also shown by Borghorst, Mulalic, and Ommeren (2021).

To conclude the discussion of model fit, I confirm in Appendix E.1 that its predicted origin-destination-level mode shares correlate strongly with data and match two other patterns. First, minibus choice probabilities decrease, and those for cars increase, with average income, just as with skill. Second, commuters substitute for poor formal public transit service with car rather than minibus trips.

FIGURE 7. COMMUTE MODE SHARES BY SKILL GROUP IN DATA VERSUS MODEL



Notes: This figure displays the shares of low- (non-college) and high-skill commuters in Cape Town who use each mode in the data and as predicted by the model. Data comes from the 2013 Cape Town Household Travel Survey; in the model, I average origin-destination choice probabilities π^g_{ijm}, weighted by inflows N^g_{ij}.

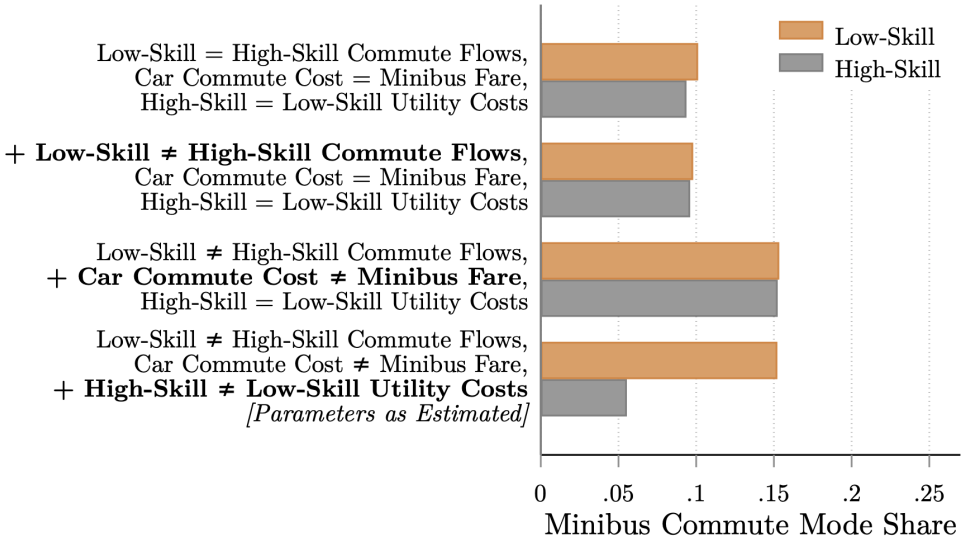
Decomposing Mode Choices

The low-skill minibus mode share in Cape Town exceeds that of the high-skill by 21 percentage points in the data. This stark skill differential could stem from (i) geography, meaning better access to formal public transit in highly-educated neighborhoods; (ii) the high monetary cost of cars; or (iii) the starkly higher minibus utility costs of the high-skill. Figure 8 demonstrates that (iii) accounts for the difference in minibus mode shares between skill groups, as follows. I begin with a model with (i) equal commute flow distributions across skill groups, $N^l_{ij} \propto N^h_{ij}$; (ii) a car cost equal to the average minibus fare, $\tau_A = \overline{\tau}_M$; and (iii) equal utility costs across groups, $\kappa^h_m = \kappa^l_m$. The top line in Figure 8 plots the resulting similar minibus mode shares across skill groups. When I reintroduce the empirical commute flows, the minibus shares in the second line scarcely change. As displayed in the third line, both skill groups shift towards minibuses when I increase the car cost to its calibrated value. Only when commuters experience the estimated utility costs of each mode, in the fourth line, do the latter abandon minibuses. Policies that reduce these costs, such as providing security, might thus offer substantial welfare gains.

VII. URBAN TRANSPORTATION POLICIES

Finally, I use the estimated model to analyze counterfactual policy strategies to optimize the existing minibus provision. I investigate (i) an optimal minibus entry tax; (ii) improvements in matching efficiency, potentially through a mobile app; (iii) the provision of security at

FIGURE 8. SKILL DIFFERENTIAL IN MINIBUS USE:
GEOGRAPHY, MONETARY COSTS, OR UTILITY COSTS?



Notes: This figure displays model-predicted skill-group-level average minibus mode shares, under different parameter assumptions that shut down alternative explanations for the skill differential in mode shares. In the first line, I equalize the low- and high-skill commute flow distributions so that $N_{ij}^l \propto N_{ij}^h$, set the car commute cost equal to the mean minibus fare, $\tau_A = \overline{\tau}_M$, and set high-skill equal to low-skill utility costs for each mode, $\kappa_m^h = \kappa_m^l$. In the second, I reintroduce the true $N_{ij}^l \neq N_{ij}^h$; in the third, the true car commute cost τ_A; and in the fourth, different utility costs across skills, $\kappa_m^h \neq \kappa_m^l$.

publicly-owned minibus stations; and (iv) the construction of exclusive minibus lanes. In terms of model objects, these policies operate on (i) the entry cost intercept $\overline{\psi}$, (ii) matching efficiency μ, (iii) utility costs κ_m^g, and (iv) road-based destination arrival rates d_{ij}.

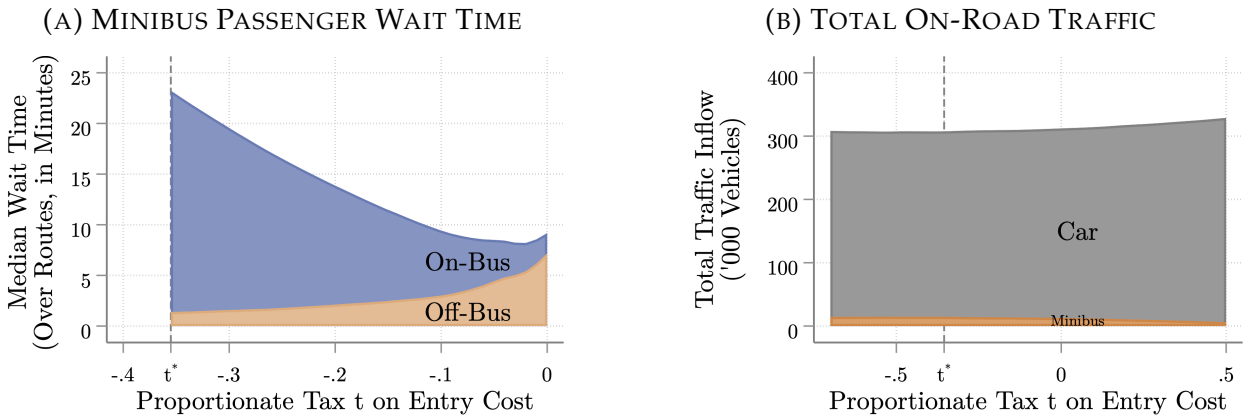
Specifically, I evaluate these policies based on utilitarian welfare. For a skill group-g commuter with home-work tuple ij, welfare $\overline{W}_{ij}^g$ equals the ex-ante expected value over commute modes, gross of any lump sum transfers T:

$$\overline{W}_{ij}^g \equiv \nu \log \left[\sum_m \exp \left[U_{ijm}^g - \kappa_m^g \right]^{1/\nu} \right] + T. \tag{23}$$

Free entry ensures that minibuses' welfare equals zero, so skill-group and aggregate utilitarian welfare correspond to commuter-inflow-weighted averages of (23).[13] I present welfare changes as equivalent variation: the proportionate change in all wages, at baseline values of $\{\boldsymbol{\lambda}, \boldsymbol{\iota}, \boldsymbol{\tau}_M, \boldsymbol{d}, \boldsymbol{\kappa}\}$, that leaves the average commuter equally well off, after tax, as in the counterfactual. Table 5 at the end of this section summarizes all counterfactuals.

[13] Taxes enter additively because utility is linear in dollars. Note that this measure equals the average welfare of a randomly chosen newly-born commuter.

FIGURE 9. MINIBUS WAIT TIMES AND ON-ROAD TRAFFIC UNDER ENTRY SUBSIDY/TAX



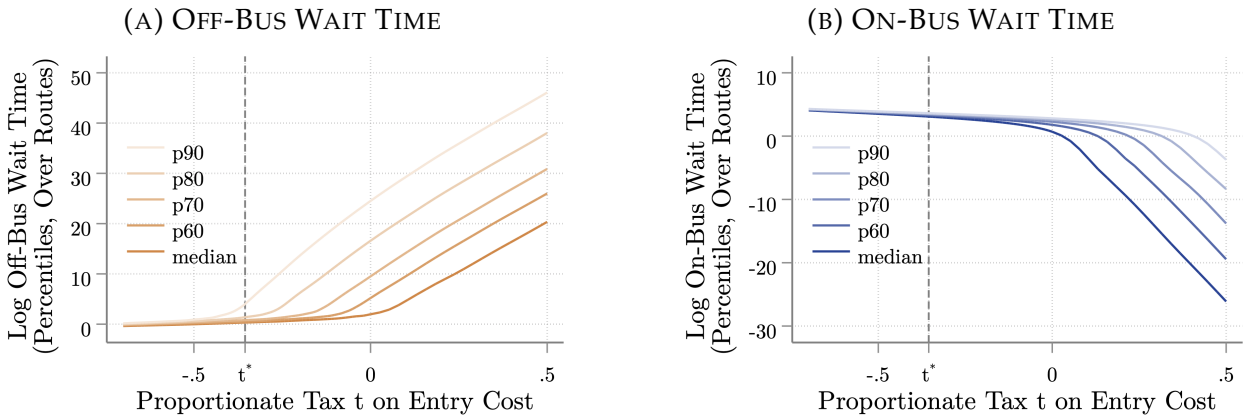
Notes: Panel (A) displays the median, across routes in the model, of expected off- and on-bus minibus passenger wait time at different levels of a proportionate tax t on minibus entry costs. Panel (B) displays the aggregate minibus plus car traffic inflow, again at different entry tax levels.

Entry Subsidy

First, I consider a minibus entry subsidy or tax to correct the boarding and filling externalities. I model the tax t as a proportionate increase in the entry cost to $(1+t)\,\overline{\psi} b_{ij}^{\phi}$, equally rebated to or financed by all commuters, in the case of a subsidy $t < 0$. I find an optimal tax of $t^* = -0.36$. Policymakers should, in fact, subsidize minibus entry because there are currently *too few* minibuses in Cape Town (see Appendix Figure A.5b). In practice, policymakers could achieve this optimal entry cost reduction through subsidies on bus purchases or direct regulation of association entry fees. My model accurately reflects the gains, provided associations do not capture the surplus in the former case and the government compensates associations in the latter case. Even after paying their share of subsidy costs, low-skill commuters gain 0.73% in welfare terms, amid a slight 0.11% loss for the high-skill.

Surprisingly, the median minibus rider does not benefit from the subsidy. Figure 9a plots a range of entry taxes on the horizontal axis and the resulting median off- and on-bus wait times across routes on the vertical axis. The optimal subsidy induces additional entry, which decreases off-bus and increases on-bus wait times relative to $t = 0$. On the median route, the latter filling externality dominates. Simultaneously, minibus fares rise slightly due to higher aggregate demand (Appendix Figure A.6d). Furthermore, as t decreases to the optimal subsidy on the horizontal axis of Figure 9b, total traffic volume dwindles because half of the new minibus riders previously drove alone (see Appendix Figures A.6b-A.6c). As a result, minibus and car travel times fall, albeit only slightly.

FIGURE 10. PERCENTILES OF MINIBUS WAIT TIME UNDER ENTRY SUBSIDY/TAX



Notes: Panels (A) and (B) display the percentiles, across routes in the model, of expected off- and on-bus minibus passenger wait time, respectively, at different levels of a proportionate tax t on minibus entry costs.

Why, then, do commuters as a whole ultimately benefit from additional entry? The losses from rising wait times on routes that already operate frequently pale compared to the welfare gains of commuters on routes where the boarding externality previously fostered extreme under-provision of minibuses. Indeed, Figures 10a-10b again display the proportionate entry tax on the horizontal axis and, on the vertical axis, the percentiles of the off- or on-bus wait distribution across minibus routes. The upper percentiles of the off-bus wait time distribution fall enormously under the optimal subsidy relative to the status quo.

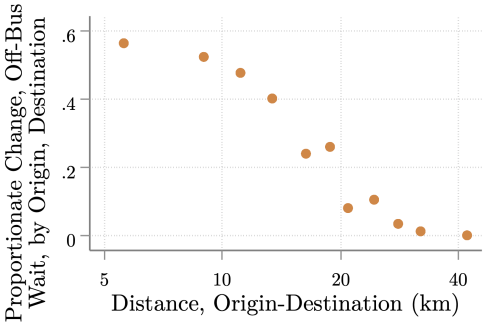
In Figure 11, I highlight the origin-destination pairs which enjoy the most substantial gains from the optimal entry subsidy. Figures 11a, 11b, and 11d display distance on the horizontal axis and, on the vertical axis, the proportionate change in a route's expected off-bus wait, the change in the minibus mode share, and the equivalent variation of commuters. Off-bus waits fall, and minibus mode shares and welfare rise, the most on origin-destination pairs slightly too long to attract sufficient bus entry without a subsidy. The routes with a greater-than-10-percentage-point increase in the minibus mode share, mapped in Figure 11c, tend to provide connections between far-flung suburbs rather than radial service to the central business district (CBD).

Matching Efficiency

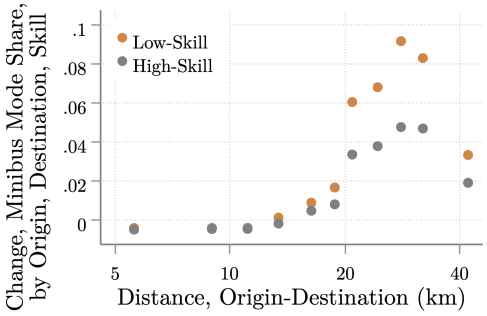
Second, the direct alleviation of matching frictions might seem a more natural solution to the wait time problem. Policymakers could realize increases in matching efficiency through

FIGURE 11. OPTIMAL SUBSIDY EFFECTS BY COMMUTE DISTANCE, VERSUS STATUS QUO

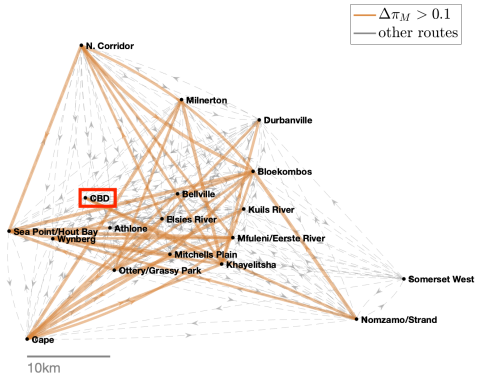
(A) OFF-(MINI)BUS WAIT TIME



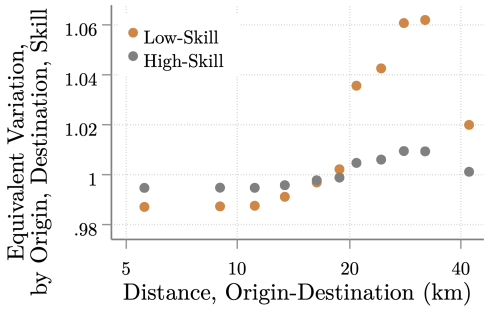
(B) MINIBUS MODE SHARE



(C) ORIGIN-DESTINATIONS WITH $> 10\%$ INCREASE IN MINIBUS MODE SHARE



(D) EQUIVALENT VARIATION



Notes: Panels (A), (B), and (D) display log-scale commute distance on the horizontal axis. On the vertical axis, they plot, respectively, the proportionate change in expected off-(mini)bus passenger wait $1/\lambda_{ij}$, the skill-origin-destination-level raw change in minibus mode share π^g_{ijM}, and the corresponding equivalent variation of commuters from implementing the optimal subsidy, all versus the zero-tax equilibrium. Panel (C) maps origin-destinations with a larger-than-10-percentage-point increase in the across-skill minibus mode share.

enlarged minibus stations or apps such as the not-yet-widely adopted Aftarobot.[14] The latter purports to directly link passengers bound for the same destination with minibus drivers, who then collect the group at an appointed time. To approximate the effects of such apps, I roughly double the baseline matching efficiency μ to the 90th percentile, $\mu' = 0.39$, of route-level matching efficiency fixed effects estimated in Column 1 of Table 2. I then solve for a new counterfactual equilibrium where minibus entry, fares, and demand adjust. Low-skill commuters gain only one-sixth as much as from the optimal entry subsidy, as evidenced in Table 5, even though I do not account for any costs of this improvement. Off-bus wait times do fall by roughly half across the board (Appendix Figure A.7a) but remain high on marginal routes. As a result, these typically longer-distance routes enjoy

[14] See http://www.aftarobot.com.

only slight minibus mode share and welfare increases (Appendix Figures A.7b-A.7c).

Minibus Security

Third, I evaluate the government provision of security guards at minibus stations. Security loomed large in my stated preference survey, and minibus operators likely under-provide guards, if they do so at all, at the publicly-owned minibus stations shared across routes. I thus adjust the minibus utility cost κ_M^g by the estimated binary security effect $\theta_{\text{security}}^g$, and commuters pay their lump-sum share of guard wages.[15] Because they place a larger premium on security, high-skill commuters shift even more strongly towards minibuses than their low-skill counterparts, as summarized in Table 5. Existing riders, however, equally enjoy the benefits of security; among these, the low-skill predominate, so low-skill welfare increases by a striking 2.5%, compared to 1.1% for the high-skill. Even within skill group, lower-income commuters switch from other modes to minibuses at higher rates and enjoy more significant welfare gains (Appendix Figures A.8a-A.8b).

Exclusive Minibus Lanes

Fourth, though minibus entry does not appreciably congest roads, high car traffic slows minibuses' trips and could contribute to significant gains from exclusive minibus lanes. I simulate the construction of such exclusive minibus lanes on the top 10% of road network links by model-predicted minibus traffic. On these links, minibuses enjoy free-flow travel times; I tax commuters to cover construction costs estimated by De Beer and Venter (2021). As a result of uniform travel time reductions across routes (Appendix Figure A.9a), minibus mode shares increase, and low-skill commuters experience net gains that slightly exceed those from the direct subsidy. Notably, on the same "marginally too long" routes that benefited from the entry subsidy, minibus use and welfare rise disproportionately (Appendix Figures A.9c- A.9d). On these routes, travel time reductions not only directly benefit commuters but also act as an indirect entry subsidy: wait times fall, albeit by less than under the direct subsidy (Appendix Figure A.9b).

Robustness

Naturally, different assumptions on entry restrictions, the level of minibus association entry fees or government red tape, minibuses' market power, or the possibility of informal

[15] The hourly guard wage, at only twice the median minibus fare, plays a minuscule role in welfare. I assume 10 guards per route to cover overhead and as a liberal, upper bound estimate; I tax commuters to cover a per-minute and guard cost quoted by a local security firm.

TABLE 5. COUNTERFACTUAL URBAN TRANSPORTATION POLICIES

		Change in Mode Share				Equivalent Variation			
		Minibus ($\Delta\pi_M$)		Car ($\Delta\pi_A$)		Pre-Tax		Post-Tax	
Policy	*Skill:*	*Low*	*High*	*Low*	*High*	*Low*	*High*	*Low*	*High*
Optimal Entry Subsidy		0.023	0.009	-0.011	-0.007	1.016	1.002	1.007	0.999
Matching Efficiency		0.002	0.001	-0.001	-0.001	1.001	1.000		
Minibus Station Security		0.032	0.039	-0.016	-0.030	1.025	1.011	1.025	1.011
Exclusive Minibus Lanes		0.018	0.006	-0.009	-0.004	1.013	1.002	1.009	1.000

Notes: This table displays the results of four counterfactuals: an optimal subsidy, $t^* = -.36$, on minibus entry costs; increasing matching efficiency to the 90th percentile of route-level matching efficiencies; adding security to all minibus stations; and building exclusive minibus lanes on the top 10% of road network segments by model-predicted minibus traffic. The first four columns show the changes in the average minibus and car choice probabilities by skill group, where I weight the origin-destination-skill-level choice probabilities π^g_{ijm} using inflows N^g_{ij}. The second four columns display skill group-level equivalent variation; the post-tax measure, for counterfactuals where I have cost data, additionally accounts for lump-sum taxation to cover the associated costs.

minibus entry might alter these conclusions. In Appendix E.3, I thus simulate alternative values of the entry cost elasticity ϕ, the entry cost intercept $\overline{\psi}$, and the fare sensitivity to demand, Γ_2. In an extension, I solve for the optimal subsidy or tax under the assumption that minibuses can enter an informal sector not subject to such taxes or subsidies. In each case, the optimal subsidy and the gains from the counterfactuals change little.

VIII. CONCLUSION

In this paper, I build a model of the privatized shared transit sector that dominates many developing-country cities. In particular, the theory highlights opposing boarding and filling externalities in minibus matching. I collected new data in Cape Town on minibus operations and user stated preferences to estimate the minibus matching function as well as the commuter demand system. Through the lens of the model, I then compare alternative counterfactual policy strategies that remedy externalities in matching, security, and road congestion.

The opposing matching externalities imply that minibus entry has ex-ante ambiguous effects on passenger wait times. In Cape Town, the positive boarding externality outweighs the negative filling externality in the aggregate. As a result, the optimal minibus entry subsidy exceeds zero and gains low-skill commuters three-fourths of a percent in welfare

terms. These average gains mask significant heterogeneity: the large decreases in off-bus wait time on marginally profitable routes more than offset limited wait time increases on the median route. The welfare effects of mobile apps that target matching efficiency pale in comparison because they fail to achieve the same benefits on long, low-demand routes. Travel rather than wait times fall upon the construction of exclusive minibus lanes on heavily-trafficked roads; welfare increases by a similar amount, net of costs, as through the subsidy. Personal safety, however, turns out to pose such a disincentive to minibus use that the gains from government-provided security guards at minibus stations dwarf those of other policies. Even without the resources to build subways or bus rapid transit, policymakers thus benefit from a range of options to improve commutes in developing-country cities like Cape Town.

REFERENCES

Allen, Teb, and Costas Arkolakis. 2022. "The Welfare Effects of Transportation Infrastructure Improvements." *The Review of Economic Studies* 0:1–47.

Almagro, Milena, Felipe Barbieri, Juan Camilo Castillo, Nathaniel Hickok, and Tobias Salz. 2022. "Public Transit Potential."

Ameriks, John, Joseph Briggs, Andrew Caplin, Matthew D. Shapiro, and Christopher Tonetti. 2020. "Long-Term-Care Utility and Late-in-Life Saving." *Journal of Political Economy* 128 (6): 2375–2451.

Andrew, Alison, and Abi Adams-Prassl. 2021. "Revealed Beliefs and the Marriage Market Return to Education."

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

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

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

Barwick, Panle Jia, Shanjun Li, Andrew R. Waxman, Jing Wu, and Tianli Xia. 2022. "Efficiency and Equity Impacts of Urban Transportation Policies with Equilibrium Sorting." NBER Working Paper 29012.

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

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

Brancaccio, Giulia, Myrto Kalouptsidi, and Theodore Papageorgiou. 2020. "Geography, transportation, and endogenous trade costs." *Econometrica* 88 (2): 657–691.

Carson, Richard T., and Mikołaj Czajkowski. 2014. *The Discrete Choice Experiment Approach to Environmental Contingent Valuation,* edited by Stephane Hess and Andrew Daly, 202–235. Cheltenham, UK: Edward Elgar Publishing.

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."

De Beer, Lourens, and Christo Venter. 2021. "Priority infrastructure for minibus-taxis: An analytical model of potential benefits and impacts." *Journal of the South African Institution of Engineering* 63 (4): 53–65.

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

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

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

Joubert, Johan W. 2013. *Gauteng: Paratransit—Perpetual Pain or Potent Potential?,* edited by Institute for Mobility Research (ifmo), 107–126. Berlin: Springer.

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

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.

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

Operating Licence Strategy 2013-2018. 2014. 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.

Pischke, Steve. 2007. *Lecture Notes on Measurement Error.*

Plano, Christopher, Roger Behrens, and Mark Zuidgeest. 2020. "Towards evening paratransit services to complement scheduled public transport in Cape Town: A driver attitudinal survey of alternative policy interventions." *Transportation Research Part A* 132:273–289.

Rayle, Lisa. 2017. "Bus rapid transit as formalization: Accessibility impacts of transport reform in Cape Town, South Africa."

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

Theway, Chesway. 2018. *Pros Cons of Minibus Taxis: The Transport System in South Africa.* https://theway.medium.com/pros-cons-of-minibus-taxis-23afd16de783.

Tsivanidis, Nick. 2019. "Evaluating the impact of urban transit infrastructure: Evidence from Bogota's Transmilenio."

Tun, Thet Hein, and Darío Hidalgo. n.d. *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.

UN. 2018. *World Urbanization Prospects: The 2018 Revision.* Technical report. United Nations, Department of Economic and Social Affairs, Population Division.

Warnes, Pablo Ernesto. 2020. "Transport Infrastructure Improvements and Spatial Sorting: Evidence from Buenos Aires."

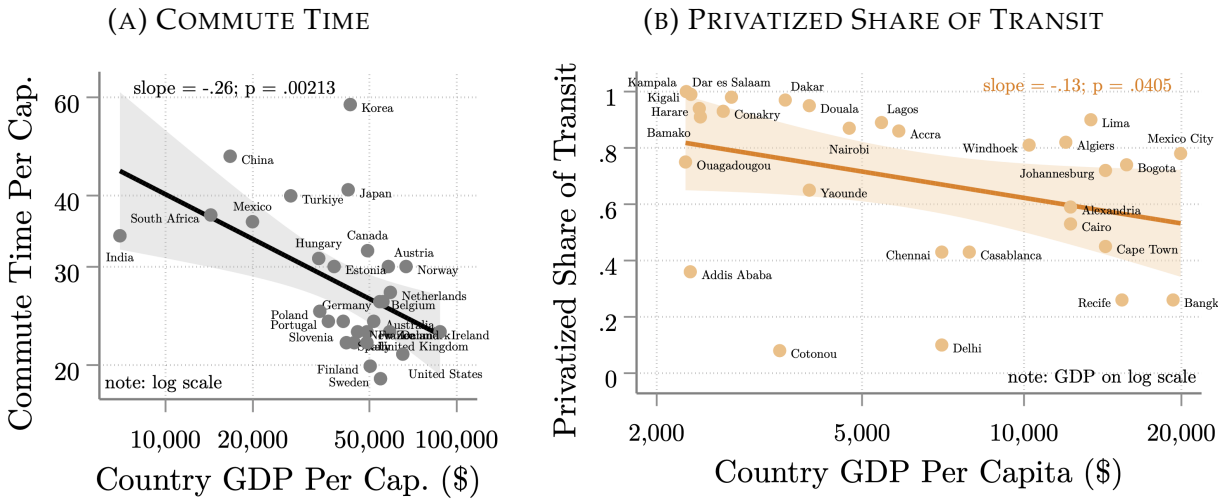
Wooldridge, Jeffrey. 2019. "The Use of Survey Weights in Regression Analysis." Presented at The Use of Test Scores in Secondary Analysis PIAAC Methodological Seminar, Paris, 14th June 2019.

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

Zarate, Roman David. 2019. "Factor Allocation, Informality and Transit Improvements: Evidence from Mexico City."

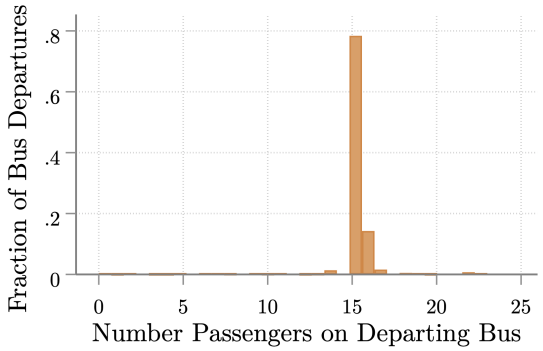
A. ADDITIONAL FIGURES

FIGURE A.1. URBAN TRANSPORTATION IN LOWER-INCOME COUNTRIES



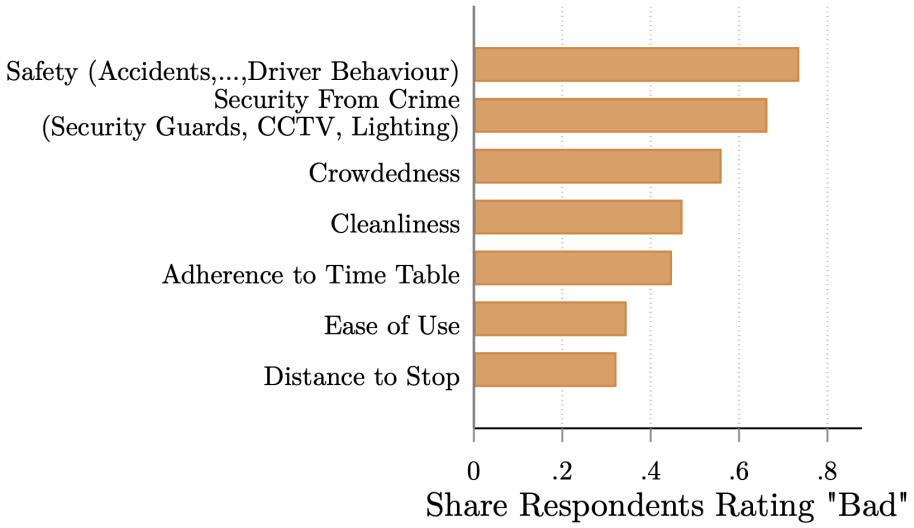
Notes: Panel (A) displays average countrywide commute time from the OECD (2016) and Panel (B) the city share of shared transit (rail, buses, and other multi-person vehicles) accounted for by private providers from Tun and Hidalgo (n.d.), both versus (national) 2019 PPP GDP per capita in current international dollars from the International Comparison Program (2020), where the units of analysis are countries and cities, respectively.

FIGURE A.2. PASSENGERS ON BOARD DEPARTING BUSES



Notes: This figure displays the distribution of the number of passengers on board departing buses, over individual bus departures on the routes in my station count data.

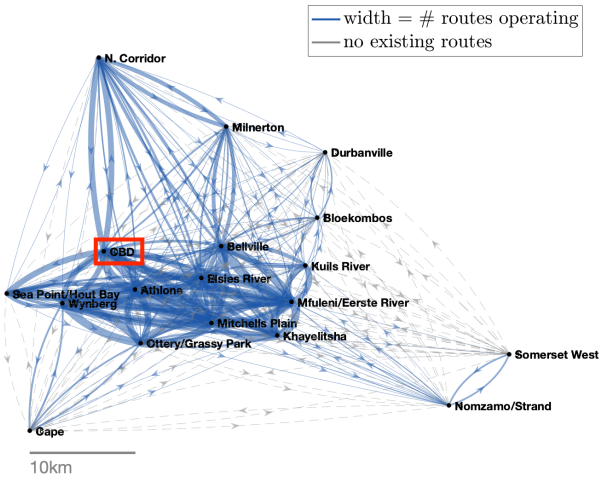
FIGURE A.3. RATINGS OF MINIBUS ATTRIBUTES IN PAST SATISFACTION SURVEY



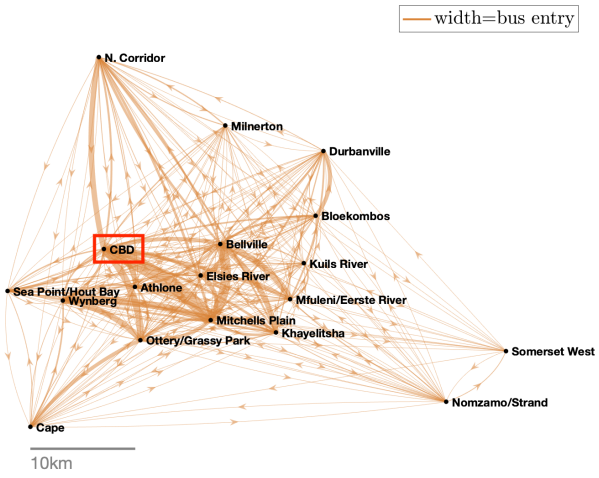
Notes: Bar graph displays share of respondents from representative sample of all mode users ($N = 1685$) in the 2013 Cape Town Household Travel Survey rating each minibus attribute as "bad," as opposed to "acceptable" or "good."

FIGURE A.4. MINIBUS NETWORK IN DATA VERSUS MODEL

(A) NUMBER ROUTES, DATA



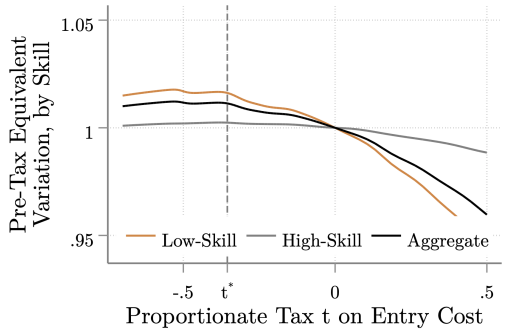
(B) MODEL-PREDICTED BUS ENTRY



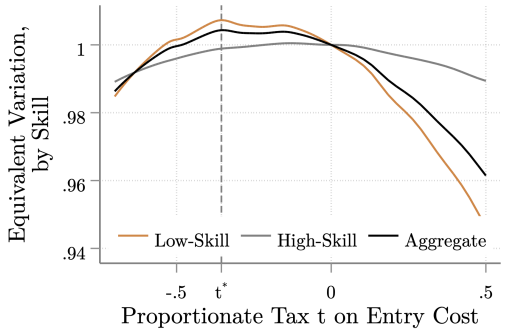
Notes: Map in Panel (A) displays number of distinct minibus routes linking each origin-destination pair of transport analysis zones according to my GoMetro network data (see Appendix B.6). Note that, since these neighborhood units include multiple minibus stations in the real world, many pairs are linked by multiple "routes" in the data, in contrast to my model. Dotted lines indicate pairs not linked by any routes. The map in Panel (B) displays origin to destination lines whose thickness corresponds to the model-predicted minibus entry per unit time on each route ij, synonymous with an origin-destination transport analysis zone tuple.

FIGURE A.5. EQUIVALENT VARIATION UNDER ENTRY SUBSIDY/TAX

(A) PRE-TAX EQUIVALENT VARIATION



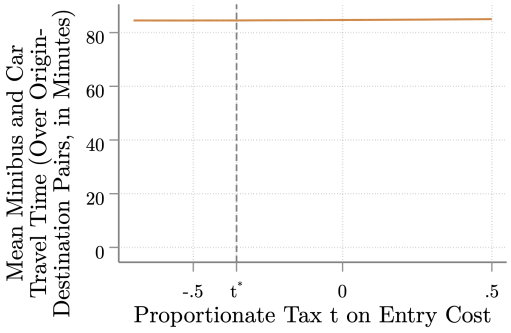
(B) POST-TAX EQUIVALENT VARIATION



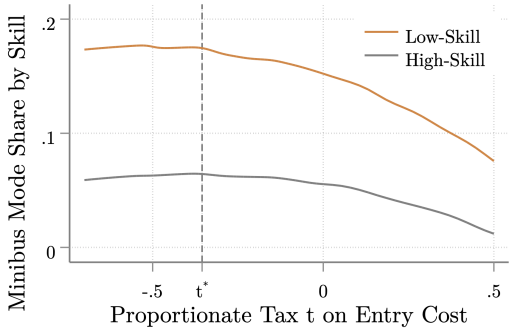
Notes: Panels (A) and (B) display equivalent variation, measured as the proportionate change Δ in wages of either a skill group or of all commuters, that would make them equally well off under the current equilibrium values of wait times, travel times, and fares as under those induced by taxing minibus entry costs at the indicated proportionate tax t. The post-tax measure in Panel (B) additionally accounts for the lump-sum redistribution of revenues or costs.

FIGURE A.6. ADDITIONAL OUTCOMES UNDER ENTRY SUBSIDY/TAX

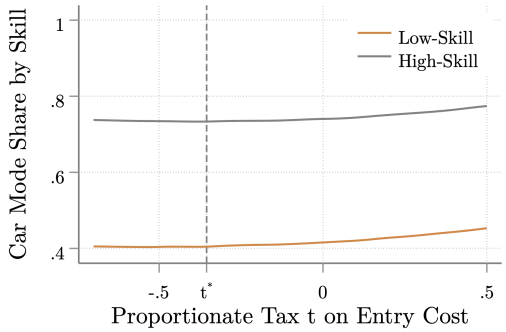
(A) MEAN MINIBUS AND CAR TRAVEL TIME



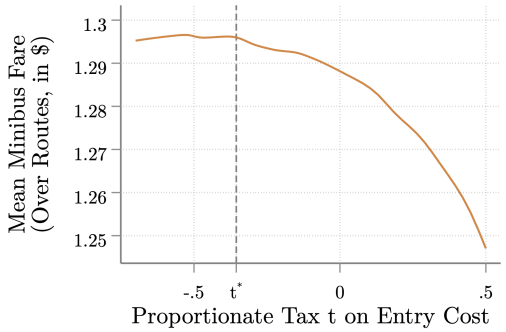
(B) MINIBUS MODE SHARE



(C) CAR MODE SHARE



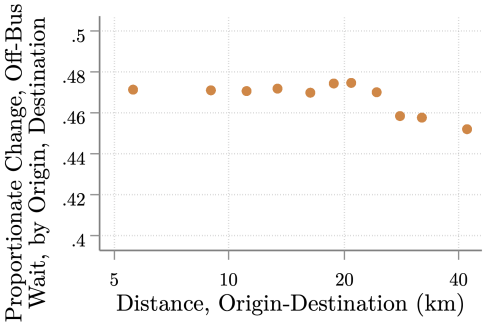
(D) MEAN MINIBUS FARE



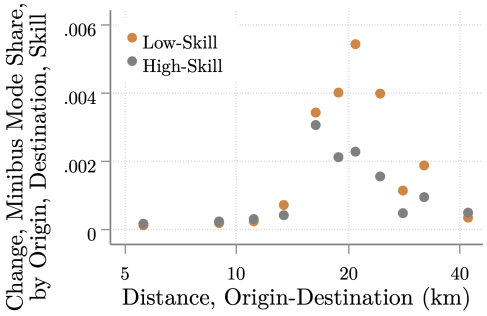
Notes: Panel (A) displays the mean, across origin-destination pairs, of expected on-road (minibus and car) travel time $1/d_{ij}$ at different levels of a proportionate tax t on minibus entry costs. Panels (B)-(C) display aggregate choice probabilities by skill, i.e. averages of the corresponding origin-destination-skill-level minibus (π^g_{ijM}) and car (π^g_{ijA}) choice probabilities weighted by inflows N^g_{ij}, and Panel (D) displays the mean minibus fare τ_{ijM} across routes (i.e. origin-destination pairs), all at different entry tax levels.

FIGURE A.7. MATCHING EFFICIENCY EFFECTS BY COMMUTE DISTANCE

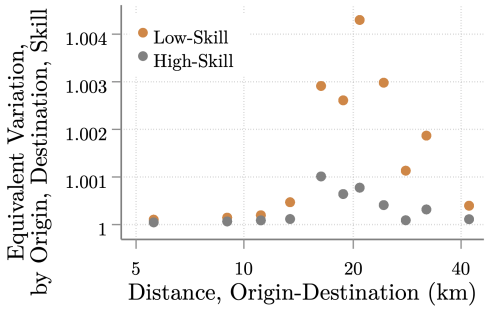
(A) OFF-(MINI)BUS WAIT TIME



(B) MINIBUS MODE SHARE



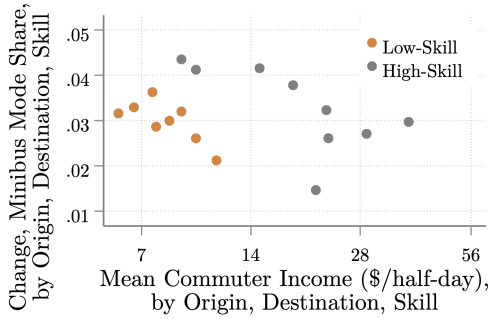
(C) EQUIVALENT VARIATION



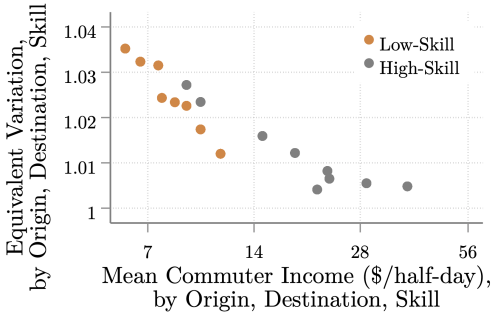
Notes: Panels (A), (B), and (C) display log-scale commute distance on the horizontal axis. On the vertical axis, they plot, respectively, the proportionate change in expected off-(mini)bus passenger wait $1/\lambda_{ij}$, the skill-origin-destination-level raw change in minibus mode share π^g_{ijM}, and the corresponding equivalent variation of commuters from increasing matching efficiency, relative to baseline.

FIGURE A.8. SECURITY EFFECTS BY INCOME

(A) MINIBUS MODE SHARE

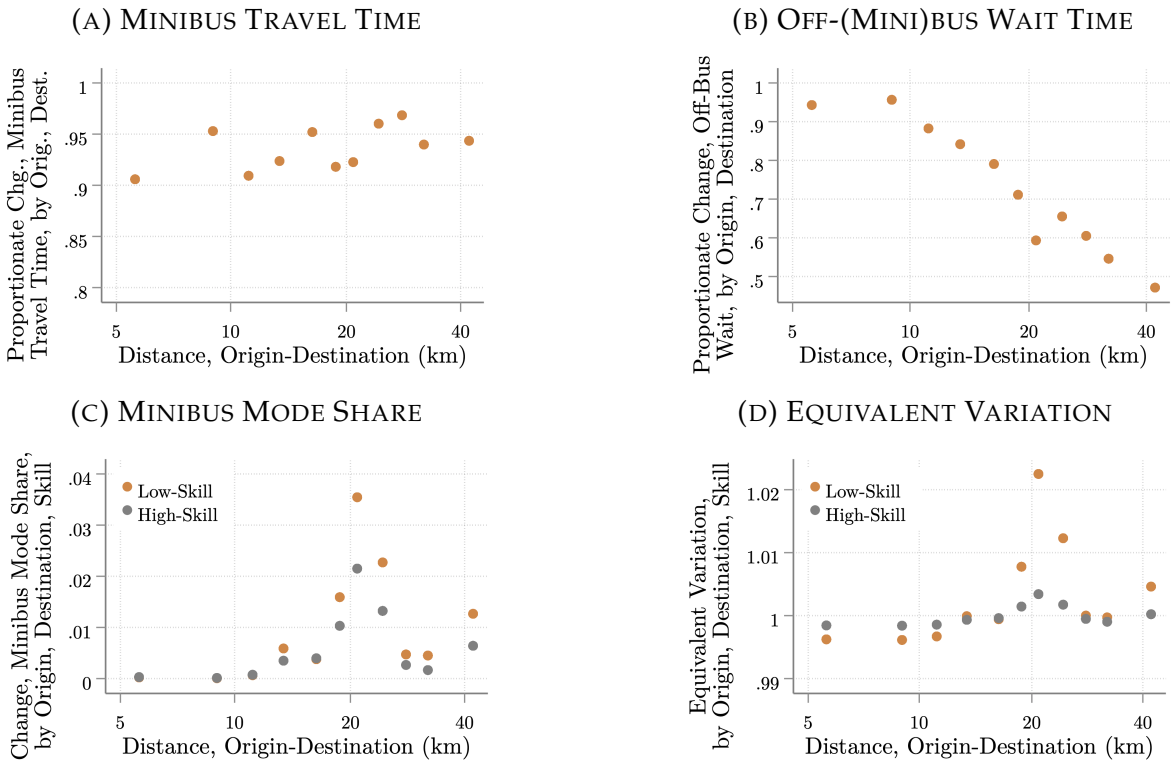


(B) EQUIVALENT VARIATION



Notes: Panels (A) and (B) display log-scale model-calibrated commuter income ω^g_j in USD per half-workday by skill, (commute) origin, and destination on the horizontal axis. On the vertical axis, they plot the skill-origin-destination-level raw change in minibus mode share π^g_{ijM} and the corresponding equivalent variation of commuters from the provision of minibus station security guards, relative to baseline.

FIGURE A.9. MINIBUS LANES EFFECTS BY COMMUTE DISTANCE



Notes: Panels (A), (B), (C), and (D) display log-scale commute distance on the horizontal axis. On the vertical axis, they plot, respectively, the proportionate change in expected minibus travel time $1/d_{ijM}$, where d_{ijM} is the arrival rate that reflects free-flow traffic; the proportionate change in expected off-(mini)bus passenger wait $1/\lambda_{ij}$; the skill-origin-destination-level raw change in minibus mode share π^{g}_{ijM}; and the corresponding equivalent variation of commuters from the exclusive minibus lane counterfactual, relative to baseline.

B. DATA APPENDIX

B.1 Minibus Station Counts

Here, I provide more detail on the minibus station counts used to characterize the matching process. I designed the counts in cooperation with the South African firm GoAscendal. GoAscendal also organized the logistics of data collection.

Scope

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 at the same station on a given day.

Sample

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. [16] This listing is, according to stakeholders, as comprehensive and accurate as any available, and includes the number minibus trips mapped by route, an albeit noisy indicator of the number of buses on a route.

Population Since, as mentioned above, the team of 12 enumerators, who could cover 6 routes, had to be employed at the same station on a given day, my survey population consists of minibus routes originating from stations in Cape Town with at least 6 routes. The aforementioned sampling frame lists 519 routes operating from 107 stations. Of these stations, 31 have at least 6 routes, and these stations account for 328 of the 519 total routes, leaving a total population of $N = 328$ routes.

Clusters and Stratification I employed a two-stage stratified cluster sample, sampling 8 stations and then 6 routes originating at each of the 8 stations. I stratify within each stage by a proxy for station and route-level bus entry 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 take 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.[17] Second, I sampled 3 stations, or a 16% sample, from the remaining 19 non-duplicate lower-bus-traffic stations with no permission issues, redrawing stations until obtaining 2 feasible ones.[18]. Figure B.1 shows the final first-stage sample of eight

[16] 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, Krogscheepers, and Spotten 2018).

[17] 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 in fact numbers 1-2, 4, 6, and 8.

[18] 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.

FIGURE B.1. SAMPLE OF MINIBUS STATIONS

Notes: This map displays the minibus station ("rank") sample at which the station counts were conducted. The five highest-bus-entry stations were sampled with 100% probability and are displayed in red; 16% of the 19 lower-bus-traffic stations were sampled and are displayed in gray.

minibus stations.

In the second stage, I take a stratified random sample of routes within each cluster, or station. Specifically, I sample 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

One station sampled, Khayelitsha (Vuyani), turned out to be part of an already sampled station (Khayelitsha (Nolungile Site C)); several others (Athlone, Vasco Station) do not operate as minibus stations with queues and loading off-road, and another set (Zevenwacht Mall, Mitchells Plain (Promenade), Tableview (Bayside)) do not operate in the AM peak.

stations.

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

B.1.1 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 lane and passenger queue corresponding to a route recorded the time a minibus vehicle arrived and also the number of passengers waiting in the queue every 5 minutes on forms such as that in Figure B.2a. The second enumerator per route monitored bus loading and departures, recording the time a vehicle began loading passengers, the time of departure, and the number of passengers on board, as in Figure B.2b.

B.1.2 Cleaning

The primary task in cleaning the data is in matching arrival times at the station for a given vehicle and route with a corresponding departure time in order to calculate the amount of time the vehicle spends waiting to load and loading each time it comes to the station. As mentioned previously, arrival times and loading/departure times were collected in separate datasets by separate enumerators. This task is not trivial not only because vehicles may arrive and depart from a station multiple times on the same day and route but also because of missing arrival and departure times. I pair vehicle arrivals with the next possible departure; when multiple arrivals occur in a row without any departures, I pair only the last arrival with the next departure.

For vehicle departures or arrivals that are not paired, I impute the corresponding missing arrival or departure. The process is simpler for vehicles that show up only once in the dataset. For these, I set the corresponding missing arrival (loading and departure) time to the beginning, or 6:00am, (end, or 10:00am), under the assumption that such vehicles

FIGURE B.2. STATION COUNT DATA COLLECTION FORMS

(A) PASSENGER QUEUES

40036 Yale Cape Town Surveys

Rank Count Form | Passenger Waiting and Vehicle Arrival

Enumerator:		Rank:		Route:		Date:	

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

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

(B) BUS LOADING AND DEPARTURE

40036 Yale Cape Town Surveys

Rank Count Form | Departure Counts

Enumerator:		Rank:		Route:		Date:	

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

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

arrive before or load and depart after the hours during which the station counts were conducted.[19] For vehicles that show up multiple times and have a non-paired departure that is not the first of the day, I average the preceding departure time and the current loading start time to obtain an imputed arrival time. If a vehicle shows up multiple times and has a non-paired arrival that is not the last of the day, I impute the loading start time and the departure time as the average of the current and next arrival time.

I perform additional cleaning by adjusting vehicle ID formats to facilitate merging, removing duplicates in terms of vehicle, route, and arrival/departure time (separately), dropping records from the departure file where the loading start time is after the departure time, and departures after 10:00. The resulting final dataset covers the 44 routes in Table B.1, distributed across 8 origin stations.

[19] I perform a similar imputation if the non-paired departure is the first of the day or the non-paired arrival is the last of the survey period.

TABLE B.1. STATION COUNT ROUTES (ORIGIN-DESTINATION)

Bellville-Bloekombos	Mfuleni-Epping
Bellville-Durbanville	Mfuleni-Killarney
Bellville-Eerste River	Mfuleni-Kuils River
Bellville-Khayelitsha Site C	Mfuleni-Wynberg
Bellville-Mfuleni	Mitchells Plain Town Centre-Beacon Valley
Bellville-Mitchells Plain Town Centre	Mitchells Plain Town Centre-Cape Town
DuNoon-Bayside	Mitchells Plain Town Centre-Century City
DuNoon-Cape Town	Mitchells Plain Town Centre-Delft
DuNoon-Century City Summer Greens	Mitchells Plain Town Centre-Tafelsig
DuNoon-Montague Gardens	Mitchells Plain Town Centre-Westgate
DuNoon-Parklands	Nomzamo-Bellville
DuNoon-Sunningdale	Nomzamo-GordonsBay
Elsies River-Belhar	Nomzamo-Somerset Mall
Elsies River-Edgemead	Nomzamo-Somerset West
Khayelitsha Site C-Claremont	Nomzamo-Stellenbosch
Khayelitsha Site C-Elsies River	Nomzamo-Strand
Khayelitsha Site C-Epping	Wesbank-Cape Town
Khayelitsha Site C-Lower Crossroads	Wesbank-Eerste River
Khayelitsha Site C-Sacks Circle	Wesbank-Kuils River
Khayelitsha Site C-Wynberg	Wesbank-Mitchells Plain Town Centre
Mfuleni-Cape Town	Wesbank-Mowbray
Mfuleni-Elsies River	Wesbank-Wynberg

Notes: This table displays the final station count sample of 44 minibus routes in Cape Town. I drew 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 bus entry. Note that Elsies River station was revealed on the survey day to have only two operating routes, resulting in a total of 44 rather than 48 routes.

B.1.3 Calculations

I discretize time into 5-minute periods, each beginning at some clock time t. I calculate the number of waiting buses on route l, b_{lt}, as the number of vehicles which arrive at the station before $t+5$ and depart after t, i.e. the total number present during that 5-minute block, based on my imputations of arrivals and departures.

Calculating the number of loading, or matched, passengers, boarding buses in a period is a bit more involved. Denote by loadingtime_{sl} the number of minutes between the time the minibus for trip s on route l begins loading and the time it departs, as described in Appendix B.1.1. I then make the non-heroic assumption that the passengers I observe departing from the origin station on trip s, deppax_{sl} board the bus at a uniform rate. Then, I apportion these passengers who depart on a trip to the five-minute blocks during which the bus loads proportionally, to calculate the total number of passengers loading (boarding)

buses, P^L_{lt} on route l from t to $t+5$:

$$P^L_{lt} \equiv \sum_s \frac{\text{loadingtime}_{sl} \cap [t, t+5)}{\text{loadingtime}_{sl}} \text{deppax}_{sl} \tag{B.1}$$

Here, $\text{loadingtime}_{sl} \cap [t, t+5)$ indicates the number of minutes of trip s's loading time that overlap temporally with clock times t to $t+5$. The number of departing passengers, P^D_{lt}, is the sum of all passengers on buses on route l which depart the station between t and $t+5$, i.e.

$$P^D_{lt} = \sum_s \mathbb{1}\left\{s \text{ departs in } [t, t+5)\right\} \text{deppax}_{sl} \tag{B.2}$$

Finally, I seek the number of unique passengers who wait during $[t, t+5)$. For each route l and period t, I observe (a static snapshot of) the number of passengers waiting exactly at t, pax wait at t_{lt}. Under the assumption that no passengers give up and stop waiting before boarding a bus, I can then calculate the total number of waiting passengers during period t, or between $[t, t+5)$, as

$$p_{lt} \equiv P^L_{lt} + \text{pax wait at t+5}_{lt} \tag{B.3}$$

The average waiting time of these passengers to board buses, consistent with the model, is

$$\text{passenger wait time}_{lt} \equiv 5 \cdot \frac{p_{lt}}{P^L_{lt}} \tag{B.4}$$

and is measured in minutes. The bus loading rate, denoted by ι in the model, is

$$\iota_{lt} \equiv \frac{P^L_{lt}}{5 \cdot b_{lt}}. \tag{B.5}$$

B.2 Minibus Stated Preference Survey

Next, I detail the stated preference survey which I conducted to estimate the demand system, also implemented by GoAscendal.

Questionnaire Design

I designed the questionnaire to maximize statistical power while retaining respondent attention. Since the Cape Town household travel survey already contains discrete choice experiments containing different modes of transport, I focus exclusively on minibus

TABLE B.2. IMPORTANCE OF MODE CHOICE FACTORS (% OF RESPONDENTS)

Rating	Access	Appearance	Comfort	Convenience	Reliab[ility]	Safety	Security
1: Not Very Important	3	3	1	9	4	1	1
2	10	6	3	6	6	6	4
3	11	16	6	11	11	10	13
4	27	27	20	26	22	14	17
5: Most Important	50	47	69	47	58	69	66

Notes: This table displays responses to the following question regarding transport mode choice, "Which of the following factors will most influence your decision? For each of the seven factors, rate them in terms of importance (1-5)", by stated preference respondents in the 2013 Cape Town Household Travel Survey ($N = 1,677$).

commutes with different non-pecuniary attributes and cost. The first step is then to choose which attributes are explicitly varied across alternatives within each choice set.

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, Hanson, and McPake 2009; Johnston et al. 2017). Conveniently, the Cape Town household survey asks respondents to rate the importance of a variety of factors in their mode choice decisions. A quick look at Table B.2 reveals that the three factors most frequently rated "most important" in mode choice are 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," as I discuss in Section II.[20]

I thus choose three nonpecuniary attributes corresponding to mode users' three main concerns: presence or absence of security guards, driver adherence to speed limits (to capture safety), 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; Mangham, Hanson, and McPake 2009), I choose attribute levels for the quantitative attributes that are plausible and within the range of typically experienced values in Cape Town yet allow for sufficient variation.[21]

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

[21] 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.

D-Efficiency Algorithm I use algorithm built into the Stata package `dcreate` to choose a *d-efficient* questionnaire design, namely the combinations of attribute levels in each alternative of each choice set presented to respondents. This algorithm minimizes the determinant of the variance-covariance matrix of the estimated parameters a given discrete choice model under some priors (Ben-Akiva, McFadden, and Train 2019; 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 $\overline{\kappa}_{cl}$ only while traveling on a minibus, rather than a one-time utility cost. I can write the utility of choice set c, minibus alternative l for individual i as

$$U_{icl} = \frac{w^{-1}}{r + w^{-1}} \left[-\frac{\overline{\kappa}_{cl}}{r + t_{cl}^{-1}} + \frac{t_{cl}^{-1}\omega_i}{r + t_{cl}^{-1}} - \tau_{cl} \right] \tag{B.6}$$

where w denotes the (constant) wait time, t_{cl} travel time, τ_{cl} the fare, ω_i personal income, and I have used the relationship between arrival rates and average travel time $t_{cl} = \frac{1}{d_{cl}}$, and a similar relationship for wait time.

I assume linear effects θ_z of each quality improvement z on the utility cost κ_M of a minibus without any quality improvements. Thus, I can write the utility cost of alternative l in choice set c, $\overline{\kappa}_{cl}$, given a dummy variable $q_{cl}(z)$ for the inclusion of a given quality improvement z in alternative l of choice set c, as

$$\overline{\kappa}_{cl} = \kappa_M + \sum_z \theta_z q_{cl}(z). \tag{B.7}$$

Finally, I take a first-order approximation to the utility in (B.6) and can then write the choice probability of alternative l in choice set c for individual i as a multinomial logit model that is linear in parameters:

$$\pi_{icl} = \overline{U}_{icl}^{-1} \exp \left[-\frac{\kappa_M + \sum_z \theta_z q_{cl}(z)}{\nu} (1 - rw)\, t_{cl} - \frac{r}{\nu}\omega_i \left(t_{cl} + w\right) - \frac{1 - rw}{\nu}\tau_{cl} \right.$$
$$\left. - \frac{r}{\nu} t_{cl}^2 \left(\kappa_M + \sum_z \theta_z q_{cl}(z) \right) \right]. \tag{B.8}$$

Coefficient Priors I now require priors on the coefficients in the model in (B.8). I obtain priors $\nu = 12.7$ $r = 0.002$, and $\kappa_M = 0.88$ from estimating a version of (B.8) with non-

constant wait time on the Cape Town household travel survey stated preference module.[22] Then, I use the results of my 1-day stated preference pilot survey (with a very similar format, $N = 20$) to estimate priors for θ_z, $z \in \{\text{security}, \text{no speeding}, \text{no overloading}\}$, obtaining $\theta_{\text{security}} = -0.18$, $\theta_{\text{no speeding}} = -0.16$, and $\theta_{\text{no overloading}} = -0.21$.[23] 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 survey for ω_i.[24]

Questionnaire Dimensions I use these priors to generate 2 "blocks," or versions, of 5 choice sets with 2 alternatives each according to the d-efficiency algorithm.[25] A sixth choice set has one strictly dominant option and is meant to test respondent attention; I do not use this sixth choice set in later estimation. I do not include an outside option (Ben-Akiva, McFadden, and Train 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, Hanson, and McPake 2009).

Additional Demographic Questions

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.

Pilot Survey Lessons

The enumerator team and I conducted a pilot survey at the Cape Town CBD minibus station on 15 June 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.

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

[23] In estimating the conditional 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.

[24] This is $9600/22.5 = 427$.

[25] 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.

Overall Good Comprehension Enumerators and I first tried to walk respondents through each choice set and attribute value verbally but soon found this tedious for all involved and discovered that one could hand the respondent the printed sheet with choice sets and allow him/her to independently read and indicate his/her choices. Anecdotally, the respondents I interviewed in this fashion 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!"

Some Fatigue and Inattention Enumerators reported respondent fatigue after approximately the 5th choice set and noted this problem seemed to be diminished in the version with only two alternatives. There was, however, no statistically significant difference in the mean time it took to complete the two versions, and all respondents except one who qualified for the survey also completed all questions. Greater cause for concern was the fact that, in the final question with a strictly dominant alternative, only 43% of those receiving 3 alternatives and 36% of those receiving 2 alternatives chose the superior option, suggesting severe deterioration of respondent attention, as this question came last. These observations, both anecdotal and statistical, motivated my simplification of the questionnaire, reducing the number of choice sets to 5 plus the attention question and the number of alternatives per set to 2.

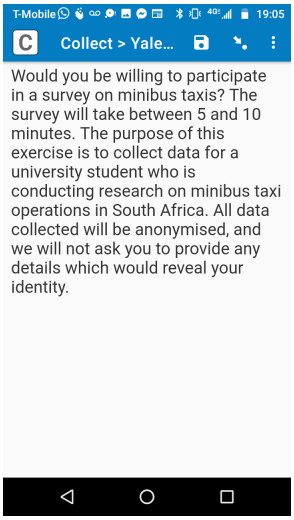
Demographic Question Modifications Additionally, lack of knowledge regarding other household members' incomes and the structure of my model motivated asking only for a respondent's own income.

Sample

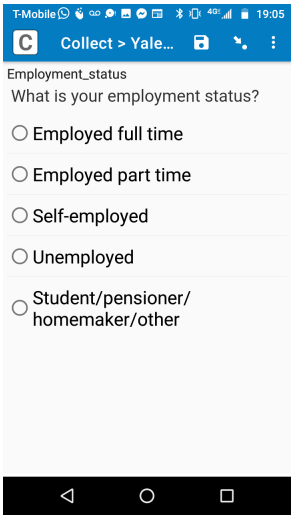
Stated preference surveys were conducted at one mall and transport interchange and at two minibus stations, at each of the three locations for 5 hours, 11am-4pm, on 21, 27, and 30 June 2022 (all weekdays). Security considerations did not permit a random sampling of minibus stations or other locations, 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 both a representative sample of different mode users (at the mall and transport interchange) and also a sample of respondents intimately familiar with minibuses, for whom the

FIGURE B.3. STATED PREFERENCE SURVEY APP SCREENSHOTS

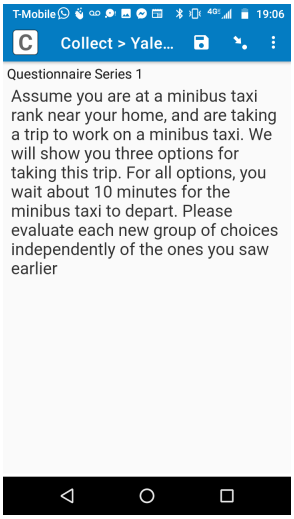
(A) CONSENT SCRIPT



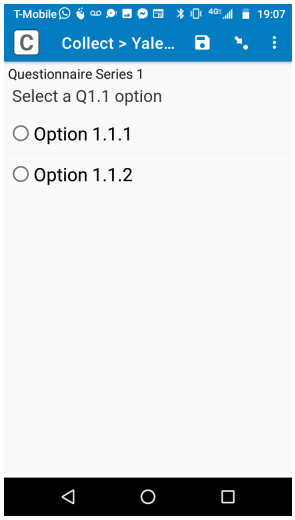
(B) DEMOGRAPHIC QUESTIONS



(C) PRE-SCENARIO SCRIPT



(D) RESPONSE ENTRY



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

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.[26]

B.2.1 Administration and Script

Survey enumerators randomly approached respondents and asked their consent to participate in the survey, according to a script in Figure B.3a. They were offered a chocolate as an incentive. Enumerators then proceeded to read them questions shown in the Survey CTO Android app (see Figure B.3b-B.3d), which automatically progresses through the questionnaire, showing follow up questions or terminating the survey where appropriate, based on previous responses. The stated preference scenarios themselves were shown on laminated paper, and enumerators also entered responses directly into the app.

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

B.2.2 Field Experience

All 526 employed people we approached completed the entire survey. The 6th choice set, common to both blocks, had a strictly dominant option and thus could provide a measure of comprehension. Approximately two-thirds of respondents chose the strictly dominant alternative. The premise of a discrete model would suggest that some share of respondents would indeed pick the strictly dominated alternative due to high idiosyncratic preference draws, but one-third a high enough share to make one doubt that all were paying attention and taking the survey seriously. Enumerators were given the opportunity to flag respondents who were distracted or disinterested, which they did for 16 percent of respondents; this flagging did not strongly predict whether respondents chose the dominant option in question 6, as 58 percent did so in the "disinterested" group and 69 percent in the non-disinterested group. Note that there were no corner solutions: every alternative of every question was chosen by a nonzero number of respondents.

B.2.3 Sample Characteristics

In later estimation, I stack my own stated preference survey with a city-conducted survey from 2013. Table B.3 compares the demographic characteristics of each sample to the population commuter population, as taken from the 2013 Cape Town Household Travel Survey. Along basic demographics, 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 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 TomTom MOVE API

The traffic data company TomTom provides a variety of traffic data drawn from locational pings of TomTom-linked GPS units and smartphone apps. The TomTom MOVE suite of products includes an API called *Traffic Stats*, whose Area Analysis feature allows the querying of historical traffic data for all road segments in a defined area and time period(s).[27]

In particular, Traffic Stats provides average travel time, average speed, the standard deviation of speed, the speed percentiles, speed limit, road name, and the functional road

[27] https://developer.tomtom.com/traffic-stats/documentation/product-information/introduction

TABLE B.3. STATED PREFERENCE SAMPLE CHARACTERISTICS

	Stated Pref. Samples		*Data*
Variable	Own	City-Run	Cape Town
Share Auto Owners	0.448	0.581	0.561
Share Female	0.458	0.494	0.458
Share College-Educated	0.295	0.228	0.190
Median Monthly Personal Income [bin]	\$182-\$364	\$182-\$364	\$182-\$364
Median Age	35	39	39
Commute Mode Shares of...			
Minibus	59.56	22.56	23.55
Formal Transit	19.61	27.69	22.81
Auto	12.11	40	39.40
Share Using Minibuses > 1x/Week	0.951	0.635	
N	413	407	

Notes: This table's first two columns display demographic characteristics of my newly-conducted stated preference sample as well as the sample used from the 2013 Cape Town Household Travel Survey stated preference module. The third lists the corresponding statistics in the aggregate Cape Town population, as calculated from 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.

class. Importantly, for my context, it also returns the number of tracked "probes," or TomTom-linked devices, which can proxy for total traffic on a link during a given time period.[28]

To obtain a dataset of manageable size, I query the traffic statistics for each road segment in the City of Cape Town, averaged over one-hour blocks from 4am to 11pm, on 13 January 2021 (a Wednesday), yielding a total of 231,707 unique road segments and $N = 2,355,901$ segment-time observations with nonzero traffic.

B.4 National Household Travel Survey (2013)

To estimate the fare supply function, I make use of the 2013 and 2020 waves of the National Household Travel Survey (NHTS), which consists of a stratified random sample of households in both urban and rural areas, corresponding to approximately 0.3 percent

[28] Assuming that the proportion of cars with TomTom devices does not vary across links, my estimation will not be affected since the specification of the congestion function uses the logarithm of traffic.

of the population. The NHTS contains information on general travel patterns outside the home for each household member as well as more detailed information about educational and work commutes such as mode, payment method, travel time, and cost. Furthermore, the NHTS partitions the country into 384 *transport analysis zones* (TAZ) to facilitate local analysis. I restrict my respondents employed outside the home with nonmissing home and work TAZs. These restrictions leave a final analysis sample of $N = 23,055$ commuters nationwide in 2013 and $N = 26,991$ in 2020.

B.5 City of Cape Town Household Travel Survey (2013)

The City of Cape Town conducted the 2013 Cape Town Household Travel Survey (CTH-HTS) on a representative sample of residents. The survey consists of three parts: a household questionnaire covering commuting and general travel behavior administered to all 22,331 households as well as detailed one-day "trip diary" records of travel patterns[29] and stated preference questionnaires from smaller but representative subsamples of participants. In addition to demographics and car ownership, this survey records, for each individual, the addresses of residence and work, along with the details of his or her commute: time departing and arriving as well as mode, payment method, and cost of each leg of the trip.

B.5.1 Spatial Unit of Analysis

The National Household Survey's Transport Analysis Zones (TAZ) partition Cape Town into only 18 units (see Figure B.4) specifically designed for transport-related statistics. Thus, I use them as the basic spatial neighborhood unit in much of my data analysis and as locations in the model. The survey comes with the text addresses of households and of workers' places of employment. To facilitate empirical analysis, I use the Google Geocoding API to link these addresses to the TAZ. After restricting the sample to respondents working outside the home and dropping those with non-geolocated or missing addresses or with missing commute mode, I arrive at a final analysis sample of $N = 17,395$ commuters resident in Cape Town.[30]

[29] Specifically, these record origin, destination, mode, and other attributes for every change of location on a chosen reference date.

[30] A total of 21,806 surveyed individuals were employed outside the home. Note that the sample includes those working outside of Cape Town, as long as their work address was successfully associated with the corresponding TAZ.

FIGURE B.4. NHTS TRANSPORT ANALYSIS ZONES (TAZ)

Notes: This map displays the 18 transport analysis zones within Cape Town used in the 2013 South African National Household Travel Survey which are the $I = 18$ locations in my estimated model.

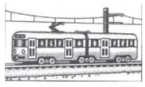
B.5.2 Defining Commute Modes

Defining mode shares or conducting analysis by commute mode is nontrivial since commuters may use multiple modes over the course of a trip. Thus, in both the NHTS and the CTHHTS, I define my mode variable based on the hierarchy below, whereby a commuter's "mode" is defined as the highest-ranked mode he or she uses over the course of the commute, where I indicate the CTHHTS-survey-defined categories that each of my aggregated mode covers in parentheses:

1. minibus (minibus/taxi)
2. formal transit (train, bus, MyCiti bus)
3. auto (car driver or passenger, motorcycle driver or passenger)
4. non-motorized or other modes (walk, bicycle, employer transport, scholar transport, other)

For example, a commuter walking to a bus stop and then taking formal transit to work would be counted as a formal rather than a non-motorized commuter. Most importantly, this means that my "minibus" category includes *anyone* who uses a minibus at some point in their commute, while "formal" includes all those who at some point use formal transit

FIGURE B.5. EXAMPLE CITY OF CAPE TOWN STATED PREFERENCE SURVEY SCENARIO

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

Notes: Each respondent in the 2013 Cape Town Household Travel Survey received "choice sets" in the format pictured, from which they indicated their preferred mode from among those listed given the associated attributes. Cost, travel time, waiting time, and number of transfers varied from choice set to choice set, and choice sets included 3 modes from among minibus taxi, train, MyCiti bus, (regular formal) bus, and car.

in their commute but do not use minibuses.[31]

B.5.3 Stated Preference Module

The stated preference module of this city-run survey corresponds to my own except that respondents choose among different *modes* of transport: car, formal MyCiti bus, (regular) formal bus, formal train, or minibus (taxi), as in Figure B.5.

B.6 Minibus On-Board Tracking Data

Third, I make use of microdata on minibus trips and stops also newly-collected by the South African firm GoMetro. This data, collected by enumerators via smartphone app, covers two trips from the beginning to the end of each route covered 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.

B.7 GoMetro Minibus Network

My final dataset maps the minibus network of Cape Town and was created through a collaboration between GoMetro and the City of Cape Town's Transport and Urban Development Authority. In order to update the city's record of on-the-ground minibus

[31] To discuss the prevalence of different modes and choices of commuters, the transportation literature generally proceeds by assigning one main mode to each commuter. Alternatively, one could count each mode used on a commute separately, but doing so would in a sense double count some commuters and be more difficult to link to my theory.

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 (Coetzee, Krogscheepers, and Spotten 2018). My dataset is a shapefile of minibus routes by day of the week and time of day, unfortunately without frequency or total passenger numbers.

Minibuses do not necessarily follow the exact same path for each trip along an official "route," complicating any attempt to rationalize trips into routes. The patterns that could be extrapolated from the GPS data collection, however, showed the official, city-designated routes to be outdated (Coetzee et al. 2018).[32] Updated route profiles were developed and recorded in the shapefile.

C. Theory Appendix

C.1 Formal Transit

Workers living in i who choose formal transit to go to work location j meet formal buses at (exogenous) rate λ_{ijF}, so that the deterministic commute utility, or value of waiting for this formal ride is

$$rU^g_{ijF} = \lambda_{ijF}\left[V^g_{ijF} - \tau_{ijF} - U^g_{ijF}\right] \tag{C.1}$$

These buses depart immediately, so that, upon boarding, group-g workers receive the value V^g_{ijF} of traveling by formal from i to j, minus the exogenous fare τ_{ijF}. The traveling value, analogously to the minibus case, reads

$$rV^g_{ijF} = d_{ijF}\left(\omega^g_j - V^g_{ijF}\right), \tag{C.2}$$

where d_{ijF} is the formal-transit-specific rate of arriving into j from i, which is not subject to congestion. Using the formal transit values (C.1)-(C.2) yields the choice probabilities for formal transit, $m = F$:

$$\pi^g_{ijF} = \exp\left[\frac{\overline{W}^g_{ij}}{\nu}\right]^{-1} \exp\left\{-\kappa^g_F + \frac{\lambda_{ijF}}{r+\lambda_{ijF}}\left[\frac{d_{ijF}\omega^g_j}{r+d_{ijF}} - \tau_{ijF}\right]\right\}^{1/\nu}. \tag{C.3}$$

[32] For example, 250 of the official city routes no longer operated (Coetzee et al. 2018)

C.2 Car

Commuters choosing car immediately pay the monetary cost of car commuting, τ_A, and depart, so that the deterministic commute utility of driving to work is simply

$$U^g_{ijA} = -\tau_A + V^g_{ijA}. \tag{C.4}$$

The corresponding traveling value V^g_{ijA} follows

$$rV^g_{ijA} = d_{ij}\left[\omega^g_j - V^g_{ijA}\right] \tag{C.5}$$

where I note that car commuters arrive at their destinations at the same road-based arrival rate d_{ij} as traveling minibus commuters. Using the Gumbel choice probabilities, (C.4), and (C.5), the probability of car commuting is given by

$$\pi^g_{ijA} = \exp\left[\frac{\overline{W}^g_{ij}}{\nu}\right]^{-1} \exp\left(-\kappa^g_A + \frac{d_{ij}}{r+d_{ij}}\omega^g_j - \tau_A\right)^{1/\nu}. \tag{C.6}$$

where $\overline{W}^g_{ij} \equiv \nu \log\left[\sum_m \exp\left[U^g_{ijm} - \kappa^g_m\right]^{1/\nu}\right]$. The flow balance equation for the mass of traveling cars is given by

$$\dot{p}^{T,g}_{ijA} = N^g_{ij}\pi^g_{ijA} - d_{ij}p^{T,g}_{ijA} = 0. \tag{C.7}$$

C.3 Steady State Flow Balance

In steady state, the number of passengers and buses in each state is constant. I now derive the corresponding flow-balance equations which help characterize equilibrium.

Minibuses First, consider the mass of loading buses on route ij with less than n passengers, $B_{ij}(n)$. Let $\varsigma_{ij}(n)$ denote the steady-state pdf of searching buses with n passengers. Because all buses start search empty and are otherwise homogeneous, the density of bus fullness levels $n \in [0, \overline{\eta}]$ is $\varsigma_{ij}(n) = \frac{1}{\overline{\eta}}$. Then, in steady state, the change in the mass of searching buses with under n passengers must equal zero:

$$\dot{B}_{ij}(n) = b^E_{ij} + (1-x)d_{ij}b^T_{ij} - \iota_{ij}\varsigma_{ij}(n)b_{ij} = b^E_{ij} + (1-x)d_{ij}b^T_{ij} - \frac{\iota_{ij}}{\overline{\eta}}b_{ij} = 0 \tag{C.8}$$

where the first term is the flow of newly entering buses in market ij. The second captures buses finishing trips on the route and not exiting. The outflow is simply the total mass

of searching buses, $b_{ij} \equiv B_{ij}(\overline{\eta})$, times the density at n times the rate at which buses meet further passengers. The mass of traveling buses from i to l, in turn, must satisfy

$$\dot{b}_{ij}^{T} = \frac{\iota_{ij}}{\overline{\eta}} b_{ij} - d_{ij} b_{ij}^{T} = 0 \tag{C.9}$$

The first term is simply the inflow of searching buses which fill up, while the subtracted outflow term captures the traveling buses arriving at j. Combining (C.8) and (C.9) yields bus entry flows:

$$b_{ij}^{E} = x d_{ij} b_{ij}^{T}. \tag{C.10}$$

Minibus Passengers I now consider minibus passengers. A flow balance equation for the number of skill-g waiting passengers on minibus route ij, p_{ij}^{g}, must hold,

$$\dot{p}_{ij}^{g} = N_{ij}^{g} \pi_{ijM}^{g} - \lambda_{ij} p_{ij}^{g} = 0 \tag{C.11}$$

where the first term captures originating inflows into the stock of waiting passengers, taking the exogenous inflow of commuters from i to final destination j and then adjusting for the share choosing minibuses. The second subtracts those who board buses. The total mass of waiting passengers relevant for the minibus matching process is then the sum across groups, $p_{ij} \equiv \sum_{g} p_{ij}^{g}$.

I next turn to passengers sitting on loading minibuses, rather than waiting at the minibus station to board. The mass of passengers sitting on loading minibuses, $\tilde{P}_{ij}^{g}$, on route ij, remains constant when the number of searching passengers meeting buses equals the mass of passengers on buses at the departure threshold $\overline{\eta}$ whose bus picks up another passenger:

$$\dot{\tilde{P}}_{ij}^{g} = \lambda_{ij} p_{ij}^{g} - \iota_{ij} \tilde{p}_{ij}^{g}(\overline{\eta}) \tilde{P}_{ij}^{g} = 0 \tag{C.12}$$

where $\tilde{p}_{ij}^{g}(n)$ denotes the pdf of bus fullness levels (masses of passengers) across skill-g *passengers* waiting on searching buses.

Finally, I consider the mass of passengers actually traveling from i to j on minibuses, $p_{ij}^{T,g}$. The inflow consists of passengers on buses with $\overline{\eta}$ passengers whose bus departs. The outflow is number of minibus passengers reaching j per unit time, $d_{ij} p_{ij}^{T,g}$. Thus, we have that, in steady state,

$$\dot{p}_{ij}^{T,g} = \iota_{ij} \tilde{p}_{ij}^{g}(\overline{\eta}) \tilde{P}_{ij}^{g} - d_{ij} p_{ij}^{T,g} = 0. \tag{C.13}$$

Formal Passengers The mass p^g_{ijF} of waiting formal passengers on route ij follows an equation analogous to that for minibus passengers:

$$\dot{p}^g_{ijF} = N^g_{ij}\left(1 - \pi^g_{ijM}\right) - \lambda_{ijF} p^g_{ijF} = 0 \tag{C.14}$$

Again following the minibus case closely, steady state flow balance for formal traveling passengers $p^{T,g}_{ijF}$ requires:

$$\dot{p}^{T,g}_{ijF} = \lambda_{ijF} p^g_{ijF} - d_{ijF} p^{T,g}_{ijF} = 0. \tag{C.15}$$

C.4 Road Congestion

In this section, I detail how the inflow of traffic using a road network link km, v_{km}, is determined. Let us first derive the inflow of traveling buses from origin i to destination j. This inflow is equal to the inflow of passengers across all skill groups g divided by $\overline{\eta}$, so that we obtain, from (C.13), bus inflow$_{ij} = \sum_g \frac{\iota_{ij}}{\overline{\eta}} \tilde{p}^g_{ij}(\overline{\eta}) \tilde{P}^g_{ij} = \sum_g \frac{d_{ij} p^{T,g}_{ij}}{\overline{\eta}}$. The inflow of traveling cars from i to j is, from (C.7) and again taking the sum over skill groups, car inflow$_{ij} = \sum_g N^g_{ij} \pi^g_{ijA} = \sum_g d_{ij} p^{T,g}_{ijA}$. In consequence, the inflow v_{km} of traffic using a link km is

$$v_{km} = \sum_i \sum_j \left[\text{bus inflow}_{ij} + \text{car inflow}_{ij}\right] \mathbb{1}\{km \in \rho(i,j)\} \tag{C.16}$$

$$= \sum_i \sum_j \left[d_{ij} \sum_g \left(\frac{p^{T,g}_{ij}}{\overline{\eta}} + p^{T,g}_{ijA} \right) \right] \mathbb{1}\{km \in \rho(i,j)\} \tag{C.17}$$

Note here that, a single minibus passenger makes only $\frac{1}{\overline{\eta}}$ the contribution to volume of a car commuter and will thus also have a smaller impact on equilibrium congestion. Finally, inserting (C.7) and traffic volume (C.16) into the definitions $t_{ik} \equiv \bar{t}_{ik} v^\gamma_{ik}$ and $d_{ij} = \left(\sum_{kk' \in \rho(i,j)} t_{kk'}\right)^{-1}$ and using the minibus flow balance equations (C.11)-(C.13), the road-based arriving-at-destination rate is

$$d_{ij} = \left[\sum_{kk' \in \rho(i,j)} \bar{t}_{kk'} \left(\sum_{i'} \sum_{j'} \left[\mathbb{1}\{kk' \in \rho(i',j')\} \sum_g \left(\frac{N^g_{i'j'} \pi^g_{i'j'M}}{\overline{\eta}} + N^g_{i'j'} \pi^g_{i'j'A} \right) \right] \right)^\gamma \right]^{-1} \tag{C.18}$$

C.5 Welfare Measure Details

To obtain skill group-level and aggregate utilitarian welfare, I average my home-work-group-specific measure $\overline{W}_{ij}^{g}$ defined in (23) across groups and home-work tuples using the exogenous inflows of workers on each relation, N_{ij}^{g}. Skill-group-level welfare, $\overline{W}^{g}$, is implicitly a function of wages, other parameters $\boldsymbol{\rho}$, and the lump-sum transfer T,

$$\overline{W}^{g}\left(\boldsymbol{\omega}^{g}, \boldsymbol{\rho}, T\right) \equiv \frac{1}{\overline{N}^{g}} \sum_{i} \sum_{j} N_{ij}^{g} \overline{W}_{ij}^{g} \tag{C.19}$$

where $\overline{N}^{g} \equiv \sum_{i} \sum_{j} N_{ij}^{g}$. Aggregate utilitarian welfare is measured as

$$\overline{W}\left(\boldsymbol{\omega}, \boldsymbol{\rho}, T\right) \equiv \frac{1}{\sum_{g} \overline{N}^{g}} \sum_{g} \sum_{i} \sum_{j} N_{ij}^{g} \overline{W}_{ij}^{g}. \tag{C.20}$$

Note that this welfare measure includes the minibuses' utility; their value is zero when taxes are zero, due to free entry, assuming that entry costs are simply lost to the economy. Under a nonzero tax, pre-tax minibus welfare is simply equal to the tax revenue and so is accounted for by adding the tax T to commuter utility in (23).

I measure changes in welfare using the *equivalent variation*, namely, the proportionate change Δ in all wages, at baseline values of $\{\boldsymbol{\lambda}, \boldsymbol{\iota}, \boldsymbol{\tau_M}, \boldsymbol{d}, \boldsymbol{\kappa}\}$, parameters $\boldsymbol{\rho}$, and zero transfer $T = 0$, that makes the average commuter equally well off, after-tax, as under counterfactual parameters $\boldsymbol{\rho'}$ and $T > 0$, i.e.

$$\overline{W}\left(\Delta\boldsymbol{\omega}, \boldsymbol{\rho}, 0\right) \equiv \overline{W}\left(\boldsymbol{\omega}, \boldsymbol{\rho'}, T\right). \tag{C.21}$$

I also calculate group- and group-home-work-location-specific equivalent variation, defined as $\overline{W}^{g}\left(\Delta^{g}\boldsymbol{\omega}^{g}, \boldsymbol{\rho}, 0\right) \equiv \overline{W}^{g}\left(\boldsymbol{\omega}^{g}, \boldsymbol{\rho'}, T\right)$ and $\overline{W}_{ij}^{g}\left(\Delta_{ij}^{g}\boldsymbol{\omega}_{j}^{g}, \boldsymbol{\rho}, 0\right) \equiv \overline{W}_{ij}^{g}\left(\boldsymbol{\omega}_{j}^{g}, \boldsymbol{\rho'}, T\right)$, respectively.

C.6 A Model of Minibus Association Entry Fee and Fare Choice

In this section, I lay out an extension of my model that incorporates minibus associations, which set entry costs as well as fares at the route level. The demand (commuter) side of the model, or the choice probabilities (9)-(11), remain identical to those in the main text, except that I reduce the model to one skill group and thus omit g superscripts. On the supply side detailed in Section IV, I shut down the road congestion channel, setting $\gamma = 0$,

for tractability; (14)-(16) continue to characterize matching and arrival rates. However, associations at the route level now choose minibus entry costs F_{ij} as well as fares τ_{ijM} to maximize the product of the mass of loading buses, b_{ij}, and per-bus expected profits, which they recoup through entry costs. Then, minibuses enter freely, subject to the chosen entry cost. The minibus problem post-entry remains unchanged.

I now discuss the association problem. To simplify the algebra, I formulate the choice of entry costs indirectly; associations first choose a mass of loading buses and a fare to maximize their objective,

$$\max_{b_{ij},\tau_{ijM}} b_{ij}\left(\overline{\eta}\tau_{ijM} - \chi\Delta_{ij}\right)\frac{1-g\left(T_{ij}\right)}{g\left(T_{ij}\right)}, \tag{C.22}$$

where the second term in (C.22) continues to characterize per-trip profits and the third, the expected number of trips that one bus makes. Associations then charge an entry cost F_{ij}^* equal to per-bus expected profits under the chosen bus loading mass b_{ij}^* and fare τ_{ijM}^*, so that under free entry, exactly b_{ij}^* loading buses enter the market with zero profits,

$$\left(\overline{\eta}\tau_{ijM}^* - \chi\Delta_{ij}\right)\exp\left[-\delta_1\left(T_{ij}\left(b_{ij}^*\right) - \delta_0\right)\right] = F_{ij}^*, \tag{C.23}$$

Note that I have indicated that the expected trip time $T_{ij}\left(\cdot\right)$ is a function of the mass of entering buses and that I have already imposed parametric assumption (20) on the bus exit rate function $g\left(\cdot\right)$.

After substituting in $T_{ij} \equiv \frac{\overline{\eta}}{\iota_{ij}} + \frac{1}{d_{ij}}$ and (15), taking first order conditions, and solving for the chosen bus entry and fares, I obtain the chosen bus mass as

$$b_{ij}^* = \frac{N_{ij}\pi_{ijM}^2}{\delta_1\overline{\eta}\left(\pi_{ijM} - b_{ij}^*\frac{\partial\pi_{ijM}}{\partial b_{ij}}\right)} \approx \frac{N_{ij}\pi_{ijM}}{\delta_1\overline{\eta}} \tag{C.24}$$

where the second approximation holds when the choice probability is not overly sensitive to the number of buses, $\frac{\partial\pi_{ijM}}{\partial b_{ij}} \approx 0$. The choice probability responds little to bus entry, holding fares constant, exactly when commuters have a low value of time, as I estimate in my stated preference data. (C.24) thus demonstrates that associations thus allow additional bus entry to adjust primarily to demand. Fares, in turn, satisfy

$$\tau_{ijM}^* = \frac{\chi\Delta_{ij}}{\overline{\eta}} - \frac{\pi_{ijM} - b_{ij}^*\frac{\partial\pi_{ijM}}{\partial b_{ij}}}{\frac{\partial\pi_{ijM}}{\partial\tau_{ijM}}} \approx \frac{\chi\Delta_{ij}}{\overline{\eta}} + \frac{\pi_{ijM}}{\left|\frac{\partial\pi_{ijM}}{\partial\tau_{ijM}}\right|}, \tag{C.25}$$

where I have used the same low-time-sensitivity approximation as well as the fact that $\frac{\partial \pi_{ijM}}{\partial \tau_{ijM}} < 0$. Fares adjust to route distance Δ_{ij} to cover buses' costs but also to demand; the latter effect grows stronger, the less lower the responsiveness $|\partial \pi_{ijM} / \partial \tau_{ijM}|$ of the minibus choice probability to fares.

Why can we say that fares increase not only with distance, but also with demand, despite the fact that associations might shift the "supply" of buses as well? As (C.24) demonstrates, bus supply does in fact increase, but this effect is swamped by another consideration. When demand levels π_{ijM} are high, buses' wait times to fill will already approach zero, so decreases in demand from higher fares are less costly, in terms of foregone trips during their finite work shift. In consequence, as long as commuters are not too time sensitive and my approximation holds, minibus fares increase with the level of demand. I employ the reduced form fare function (22) in the main text exactly to approximate the fact that minibus fares increase with distance and demand, a fact which the data bears out. I estimate a general log-linear function rather than (C.25) to additionally capture the political constraints on fare-setting discussed in Section II.

The entry costs F_{ij}^* necessary to induce entry of b_{ij}^* buses then, from (C.23), are approximately

$$F_{ij}^* \approx \frac{\overline{\eta} \pi_{ijM}}{|\frac{\partial \pi_{ijM}}{\partial \tau_{ijM}}|} \exp\left[-\delta_1 \left(T_{ij} - \delta_0\right)\right] = \frac{\overline{\eta} \iota_{ij} b_{ij}^*}{N_{ij} |\frac{\partial \pi_{ijM}}{\partial \tau_{ijM}}|} \exp\left[-\delta_1 \left(\frac{\overline{\eta}}{\iota_{ij}} + \frac{1}{d_{ij}} - \delta_0\right)\right]. \tag{C.26}$$

where the second equality uses (15) and substitutes for T_{ij}. As long as the bus loading rate ι_{ij} and the fare sensitivity of demand $|\partial \pi_{ijM} / \partial \tau_{ijM}|$ do not differ too substantially across routes, then, the entry costs charged by minibus associations increase with the mass of loading buses "chosen" by the association. Greater demand allows associations to charge higher fares without waiting forever to fill, which they can in turn recoup through higher entry cost. The entry cost congestion formulation in the main text, $F_{ij} \equiv \overline{\psi} b_{ij}^{\phi}$, can thus be viewed as an approximation of the combined effects associations' decisions and of other exogenous congestion effects, such as increases in the price of auxiliary services for minibus owners.

D. Estimation Appendix

D.1 Matching Function

I conduct two robustness checks on the estimation of the matching function. First, I reweight my regressions by inverse sampling probabilities to account for any correlation between the stratification variable and the bus loading rate. Second, I discuss measurement error issues: how my instrument might address them and also an alternative specification for the number of matches that does not rely on dividing the left-hand side variable by a right-hand side variable.

D.1.1 Sample Stratification: Weighted Estimates

I stratified each stage of my sample by the number of trips mapped on a minibus route in a previous 2018 city-commissioned on-board tracking study, which discussions with stakeholders revealed to be a measure of bus traffic on a route. In my model, then, this empirical measure should most closely proxy the number of loading buses b_{ij} that appears on the right-hand side of estimating equation (17). Under common assumptions on error term exogeneity, the unweighted estimates using a sample stratified on an independent variable will be consistent (Wooldridge 2019). However, one might be concerned that this past-trips-mapped measure might also be correlated with my outcome, namely the speed at which buses fill up, or loading rate. In such a case, the estimates in Table 2 would be biased. To explore this possibility, I reestimate all specifications, in Table D.1, while weighting each observation by the inverse sampling probability of the route with which it is associated. These probabilities are the product of the associated origin station's first-stage sampling probability and the route's second-stage sampling probability within that origin. Reassuringly, the estimates, in particular those in Column 3, which correspond to those I use in the calibrated model, change little.

D.1.2 Measurement Error

Second, I address the concern that my baseline bus loading rate specification (17) might be susceptible to measurement error because the loading rate is in fact a function of an independent variable, namely the number of loading buses b_{ij}. I first discuss the potential for my instrument to purge any bias and then estimate an alternative matching equation.

Instrumental Variables Solution I follow the framework in Pischke (2007) very closely to explore the implications of measurement error in the number of searching buses. Suppose

TABLE D.1. WEIGHTED MATCHING FUNCTION ESTIMATES

	OLS			*IV* with $\alpha + \beta = 1$	
	(1)	(2)	(3)	(4)	(5)
Parameter	log bus loading rate	log bus loading rate	log bus loading rate	log bus loading rate	log bus loading rate
α	0.703***	0.677***	0.651***	1.000***	1.348
	(0.0148)	(0.0249)	(0.0309)	(0.281)	(39.88)
β	0.405***	0.393***	0.403***		
	(0.049)	(0.053)	(0.045)		
95% CI for $\alpha + \beta$	[0.99,1.22]	[0.94,1.19]	[0.94,1.17]		
Route FE	✓	✓	✓	✓	✓
Time FE		✓	✓		✓
Origin-Time FE			✓		
Observations	1,627	1,627	1,607	1,316	1,316
R-Squared	0.600	0.846	0.861	0.45	0.0759
First-Stage F Statistic				6.20	0.00

Notes: Robust standard errors in parentheses, clustered at the origin level; *** indicates p<0.01, ** p<0.05, and * p<0.1. Each specification is weighted by the route-level inverse sampling probabilities yielded by my two-stage stratified cluster design. The unit of analysis is a minibus route (defined by origin and destination, $n = 44$) by five-minute time block in my station count data over the course of the 6-10am morning commute. Columns 1-3 present estimates of (17), with fixed effects included, as noted, for route, time (5-minute clock time block), or origin station (of the route) by time. In Columns 4-5, I assume constant returns to scale, $\beta = 1 - \alpha$, so that I can regress the log bus loading rate on the log ratio of the stock of waiting passengers to the stock of loading buses. I instrument for this ratio of passengers to buses using the log number of commuters living in the mesozone where the route originates who report leaving their home during the 15-minute period including time t, calculated from the 2013 Cape Town Household Travel Survey.

I am interested in the effect of some independent variable x on an outcome y, related according to

$$y = x + \varepsilon, \tag{D.1}$$

where ε is an error term such that $\text{cov}(x, \varepsilon) = 0$. Suppose further that y is in fact a function of some other outcome w as well as x according to $y = w - \zeta x$. In my case, w corresponds to the total loading passengers per minute, and x to the number of loading buses, the ratio of which (or difference in logs) gives the bus loading rate. I observe w and x with measurement errors ν and u, respectively, i.e. I observe $\tilde{w}$ and $\tilde{x}$ such that

$$\tilde{w} \equiv w + \nu \tag{D.2}$$

$$\tilde{x} \equiv x + u \tag{D.3}$$

Now, I substitute (D.2) into (D.1) to obtain the equation which I estimate,

$$\tilde{y} \equiv \tilde{w} - \zeta\tilde{x} = \gamma\tilde{x} \underbrace{-(\zeta+\gamma)u + \nu + \varepsilon}_{\equiv\epsilon}. \tag{D.4}$$

The composite error term ϵ thus depends on both sources of measurement error. Furthermore, measurement error in the number of loading buses in my station counts that is correlated with the true number of loading buses would bias my estimates even if the left-hand side did not depend on x, i.e. $\zeta = 0$. This bias might be magnified by a $\zeta \neq 0$, however.

My instrument addresses measurement error both in x and w as long as it is uncorrelated with the measurement errors u and ν. Were I to instrument for $\tilde{x}$ with some variable x, the instrumental variables estimator equals

$$\hat{\gamma}^{IV} = \frac{\gamma\text{cov}(x,z) + \text{cov}(\varepsilon,z) - (1+\gamma)\text{cov}(u,z) + \text{cov}(\nu,z) + \text{cov}(\varepsilon,z)}{\text{cov}(x,z) + \text{cov}(u,z)} \tag{D.5}$$

which will be consistent if and only if all terms other than (x,z) equal zero. Thus, when the instrument is uncorrelated with both u and ν, it will purge measurement error even if the *same* measurement error effects the left- and righ-hand side variables. Since my instrument is measured in 2013 and the station count data comes from 2022, it would seem highly unlikely that my instrument is correlated with any measurement error in bus loading rates or loading buses.

Alternative Specification: Equation for "Matches" Nevertheless, to further build confidence that employing a dependent variable that is a function of one of the independent variables does not inflate my measurement error bias, I re-estimate equivalent specifications for match_{ijt}, or the number of total passengers ("matches") boarding buses per minute on route ij at time t, which is not a function of either loading busees b_{ij} or passengers p_{ij}. I have only one instrument, and the total matches depend on both passengers and buses–not only on the ratio of the two, even under constant returns. Thus, I cannot estimate IV specifications for match_{ijt}. I do, however, estimate

$$\log \text{match}_{ijt} = \alpha \log p_{ijt} + \beta \log b_{ijt} + \overline{\mu}_{ij} + \overline{\mu}_t + \overline{\mu}_{it} + \varepsilon_{ijt}. \tag{D.6}$$

by OLS with the various combinations of fixed effects; Table D.2 shows that the estimated matching elasticities do not change. Thus, any additional bias from estimating an equation for the bus loading rate, rather than total boarding passengers, or matches, must be small.

TABLE D.2. MATCHING FUNCTION ESTIMATES: MATCHES AS OUTCOME

	OLS		
	(1)	(2)	(3)
Parameter	log passengers boarding/min.	log passengers boarding/min.	log passengers boarding/min.
α	0.703***	0.677***	0.651***
	(0.0148)	(0.0249)	(0.0309)
β	0.405***	0.393***	0.403***
	(0.0490)	(0.0526)	(0.0452)
Route FE	✓	✓	✓
Time FE		✓	✓
Origin-Time FE			✓
Observations	1,627	1,627	1,607
R-Squared	0.724	0.902	0.915

Notes: Robust standard errors in parentheses, clustered at the origin level; *** indicates p<0.01, ** p<0.05, and * p<0.1. Each specification is weighted by the route-level inverse sampling probabilities yielded by my two-stage stratified cluster design. The unit of analysis is a minibus route (defined by origin and destination, $n = 44$) by five-minute time block in my station count data over the course of the 6-10am morning commute. Columns 1-3 present estimates of (D.6), with fixed effects included, as noted, for route, time (5-minute clock time block), or origin station (of the route) by time.

D.2 Stated Preference

D.2.1 Implementation

Here, I provide additional details on the implementation of the stated preference estimation. I restrict the sample to individuals employed outside the home between 25 and 65 years old. Personal income ω_i is directly collected in my survey and imputed from the household section of the household survey. To calculate daily personal income ω_i in the Cape Town survey, I take the midpoint of the household's income bin, divided by 22.5 (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 one commute. For my own survey, I make similar adjustments, except that I have personal income directly instead of needing to impute it from household income. Finally, I convert all monetary amounts, including also fares, to USD for scaling purposes.

To estimate a multinomial logit, I rewrite first-order approximation in (12) to correspond to the stated preference context: the total utility u^g_{icl} for individual i with skill g of alternative l in choice set c then depends on the total mode-specific utility cost (the first term in

parentheses) as well as the indicated wait and travel times, w_{cl} and t_{cl} as well as fares τ_{cl}:

$$u_{icl}^{g} \equiv -\left(\kappa_{m(c,l)}^{g} + \sum_{z} \theta_{z}^{g} q_{cl}(z)\right) - r\omega_i \left(w_{cl} + t_{cl}\right) - \tau_{cl} + rw_{cl}\tau_{cl} + \zeta_i + \epsilon_{icl}^{g}. \tag{D.7}$$

I estimate the parameters using only the first three terms' coefficients. Each alternative is associated with a bundle of quality improvements indexed by z, where the dummy variables $q_{cl}(z)$ indicate the presence of an improvement in alternative l of choice set c. I assume that mode-specific utility cost, i.e. the first term, depends linearly on the presence or absence of each quality improvement: $\kappa_{cl}^{g} = \kappa_{m(c,l)}^{g} + \sum_z \theta_z^g q_{cl}(z)$, where $m(c,l)$ denotes the mode of transport of alternative l in choice set c. The Cape Town stated preference survey does not specify quality improvements, so $q_{cl}(z) = 0$ for all formal transit (train, (formal) bus, and MyCiti bus) and car alternatives. I thus estimate a single utility cost for each, κ_F^g and $\kappa_A^g \equiv 0$, where I normalize the latter to zero since I cannot identify the absolute level of utility costs. ζ_i denotes an individual fixed effect, and ϵ_{icl}^{g} the Gumbel-distributed preference shock.

Note further that I have used the relationship between arrival rates and average travel time $t_{cl} = \frac{1}{d_{cl}}$, and a similar relationship for wait time. My survey states that the minibus will take respondents straight to work, while the Cape Town survey additionally specifies a number of transfers to be made, presumably between vehicles of the same mode. The survey does not specify how much of each time and fare is, for example, spent on the first versus the second vehicle of a one-transfer trip; to link these scenarios to the quantitative model, I assume that the wait, travel time, and fare of all segments but the last approach 0, so that $\lambda \to \infty$ and $d \to \infty$ for, say, the first two vehicles ridden on a three-segment trip.

D.2.2 Robustness

I demonstrated in Appendix B.2.3 that my new stated preference oversamples minibus users. I test for bias resulting from this non-representative sample in two fashions. First, I reestimate the model using the city-conducted survey plus those respondents in my survey interviewed at the Middestad Mall/Bellville intermodal, a recruitment location less prone to oversampling of minibus riders. Column 2 of 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 shape ν, and the high value the high-skill place on minibus station security.

Even this "intermodal sample," however, still oversamples minibus commuters, in Column

TABLE D.3. STATED PREFERENCE: ROBUSTNESS TO SAMPLE

Parameter	Skill	(1) Baseline	(2) Intermodal Sample Only	(3) Commute Mode-Weighted
r		0.001***	0.0014**	0.0011**
		(0.0004)	(0.0007)	(.0005)
ν		4.76***	6.83**	5.84***
		(1.26)	(2.73)	(1.99)
κ_M	*Low*	7.68***	10.61***	9.25***
		(1.56)	(3.54)	(2.55)
	High	15.03***	21.16***	18.3***
		(3.55)	(7.82)	(5.67)
θ_{security}	*Low*	-1.09***	-2.13**	-1.55**
		(0.39)	(1.06)	(0.69)
	High	-2.75***	-4.91**	-5.1***
		(0.84)	(2.29)	(1.86)
$\theta_{\text{no overloading}}$	*Low*	-1.38***	-2.02**	-1.26**
		(0.437)	(1.01)	(0.596)
	High	-1.39**	-1.25	-1.43*
		(0.543)	(1.28)	(0.83)
$\theta_{\text{no speeding}}$	*Low*	-1.36***	-3.03**	-2.12**
		(0.44)	(1.38)	(0.85)
	High	-0.825*	-1.86	-0.582
		(0.465)	(1.39)	(0.73)
κ_F	*Low*	3.63***	4.53***	4.14***
		(0.51)	(1.08)	(0.80)
	High	9.17***	12.5***	10.96***
		(1.89)	(4.20)	(3.05)
N Respondents		820	546	820

Notes: Robust standard errors in parentheses; *** indicates p<0.01, ** p<0.05, and * p<0.1. 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 (19). 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.

3, I weight my own sample 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. This reweighting does not correspond to "weighting by the outcome": these respondents' choices do *not* identify the the relative mode amenity terms κ_M and

TABLE D.4. ESTIMATION OF ENTRY COST ELASTICITY ϕ

Variable	(1) mean bus loading time
log loading buses, b_{ijt}	-1.434**
	(0.546)
Constant	7.287***
	(0.332)
Route FE	✓
Origin-Time FE	✓
Observations	1,075
R-Squared	0.654

Notes: Robust standard errors in parentheses, clustered at the origin level; *** indicates p<0.01, ** p<0.05, and * p<0.1. The unit of analysis is a minibus route (defined by origin and destination, $n = 44$) by five-minute time block over the course of the 6-10am morning commute. Each column presents coefficients from an OLS regression of the bus loading time, calculated as the minutes spent loading at the station where the route originates, on the log stock of loading buses for that route at the origin station, all calculated from newly-collected station count data from Cape Town. The specification also includes fixed effects for route and origin-time (origin station of a route by 5-minute period).

κ_F since my survey asked them to choose only among *minibus* options with different attributes. Note also that the in-sample shares I use for weighting correspond to the mode the respondent typically uses to commute to work, not the mode he or she used for the "trip" before or after which the interview occurred. Reassuringly, the key takeways remain, including a low value of time and cost sensitivity as well as the high minibus utility costs and value of security.

D.3 Minibus Entry

I quantify elasticity ϕ of entry costs with respect to minibus entry with the same station count data used in the matching section. Free entry requires that bus loading times on route ij at time t decrease in the number of loading buses b_{ijt}: $\frac{\overline{\eta}}{\iota_{ijt}} = \zeta_0 - \frac{\phi}{\delta_1} \log b_{ijt} + \zeta_{ij} + \zeta_{it} + \varepsilon_{ijt}$, an equation which I estimate in Table D.4.

D.4 Minibus Fares

To estimate the dependence of fares on distance, i.e. the coefficient Γ_1 in (22), I use the onboard minibus tracking data and calculate the mean fare $\overline{\tau}_{ijM}$ and straight-line distance, $\overline{\Delta}_{ij}$, from origin to destination of a given route ij. I then estimate $\log \overline{\tau}_{ijM} = \tilde{\Gamma}_0 + \Gamma_1 \log \overline{\Delta}_{ij} + \epsilon_{ij}$ by OLS and report the positive fare-distance gradient in Table D.5. The

TABLE D.5. ESTIMATION OF MINIBUS FARE DISTANCE SLOPE Γ_1

Variables	(1) log mean fare
log straight-line distance	0.292***
	(0.0232)
Constant	2.231***
	(0.0591)
Observations	43
R-Squared	0.798

Notes: Robust standard errors in parentheses; *** indicates p<0.01, ** p<0.05, and * p<0.1. The unit of analysis is a minibus route (defined by origin and destination, $n = 44$) in my on-board tracking data. Each column presents coefficients from an OLS regression of the (log) mean fare paid on a route on the log average straight-line distance from origin to destination of trips on that route as well as a constant.

TABLE D.6. ESTIMATION OF MINIBUS FARE DEMAND SLOPE Γ_2

Variable	(1) *OLS* log mean fare	(2) *OLS* log mean fare	(3) *IV* log mean fare	(4) *IV* log mean fare
log minibus commuters	0.0780***	0.0592***	0.0438**	0.0273
	(0.0148)	(0.0204)	(0.0191)	(0.0255)
Year FE	✓	✓	✓	✓
Municipality FE		✓		✓
Observations	649	614	649	614
R-Squared	0.189	0.551	0.180	0.021

Notes: Robust standard errors in parentheses; *** indicates p<0.01, ** p<0.05, and * p<0.1. The unit of analysis is a transport analysis zone (TAZ) in South Africa by year (2013 or 2020). Each column presents coefficients from a regression of the log mean minibus fare in paid in a TAZ on the log number of residents commuting to work by minibus as well as a constant and time or municipality fixed effects, as noted. In Columns 3-4, I instrument for the number of minibus commuters with the (log) total number of employed residents.

high R^2 confirms distance as fares' primary determinant.

I estimate Γ_2, the effect of citywide minibus demand on fares, from the 2013 and 2020 waves of the South African National Household Travel Survey. For each of 384 transport analysis zones across South Africa, which are smaller than a municipality but larger than a census tract, I calculate the mean fare paid by minibus commuters as well as the total number of residents commuting by minibus and regress the former on the latter in Table D.6. The microfoundation which this regression approximates implies that the primary

threat to identification comes from supply shifters: variations in operating or entry costs across zones. I first by including municipality fixed effects in Column 2 to account for operating costs such as gas unlikely to vary within metropolitan regions. Then, in Columns 3-4, I instrument for minibus commuters with total population, a valid instrument if larger cities do not have higher operating costs, conditional on the size of the minibus market. I find a low supply slope, Γ_2: a doubling of demand would increase fares by only 6%. The slope changes little until the instrumented fixed effects specification in Column 4, where precision falls.

In the model, I use the Column 3 estimate, $\Gamma_2 = .0438$, and set $\Gamma_0 = \tilde{\Gamma}_0 - \Gamma_2 \log\left(\sum_{i,j,g} N_{ij}^g \pi_{ijM}^g\right)$, where the latter term reflects the baseline equilibrium total minibus demand calculated by setting $\Gamma_0 = \tilde{\Gamma}_0$ and $\Gamma_2 = 0$. As a result, the baseline minibus fares τ_{ijM} equal the predicted values from Table D.5 using the straight-line distance between the (emoloyed population-weighted centroids of) locations i and j. Thus, the fares in the baseline equilibrium are essentially exogenously calibrated.

D.5 Externally Calibrated Parameters

D.5.1 Road Congestion

Table D.7 displays the results from a regression, in the TomTom MOVE 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.[33]

D.5.2 Commute Flows and Wages

I calculate commute flows N_{ij}^g between transport analysis zones in the City of Cape Town using the 2013 Cape Town Household Travel Survey (see Appendix B.5) and the accompanying sample weights, excluding those who work in the same transport analysis zone. I define two skill groups g, high and low, where high-skill includes those with a tertiary degree.

I also use the 2013 Cape Town survey to compute wages ω_j^g by skill group and transport analysis zones, taking the weighted mean daily per-person household income of workers in a skill group employed outside the home whose workplace is located in a TAZ, regardless of where they live.[34]

[33] Motorway designates functional road class 0, major/secondary classes 1-3, major local classes 4-5, and minor classes 6-7.

[34] I divide monthly income by 22.5, the average number of working days in a month, to obtain daily

TABLE D.7. ESTIMATION OF ROAD CONGESTION ELASTICITY γ

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

Notes: Robust standard errors in parentheses; *** indicates p<0.01, ** p<0.05, and * p<0.1. 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. Both specifications include segment fixed effects.

D.5.3 Free-Flow Driving Times and Distance

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.[35] To approximate the no-traffic travel times, I query predicted travel times for Sunday, May 8, 2022 at 11pm. The driving distance Δ_{ij} that determines minibuses' operating costs equals the distance driven along the route chosen in these queries.

income, divide by 2 since I only model a one-way commute, and additionally convert income to USD for scaling purposes.

[35] I calculate the center of employment as the weighted average centroid latitude and longitude of *small area layer* (SAL) units in the TAZ, where the weights are employment by residence in the SAL times the proportion of SAL area within the TAZ. Note that I also use only the part of the SAL unit within the TAZ to calculate centroids. SAL is the smallest census geography.

D.5.4 *Formal Transit Arrival Rates λ_{ijF} and d_{ijF}*

Using the Microsoft Azure API (similar to Google Maps), I calculate travel times between centers of employed population of transport analysis zones (TAZ) in Cape Town on formal transit, including Metrorail commuter trains, Golden Arrow private scheduled buses, and MyCiti government-run buses. I do so by querying 6 evenly-spaced trip times on a Wednesday between 7:00 and 8:00am, and take the average waiting time for the first transport vehicle of a trip and average travel time after boarding the first vehicle.[36] Then, using properties of Poisson arrival processes, the arrival rate of formal transit to a stop, λ_{ijF}, is calibrated as the inverse of the average wait time.[37] The formal transit arrival at destination rate d_{ijF} is calibrated as the inverse of the average travel time between TAZ i and TAZ j.

D.5.5 *Formal Transit Fares τ_{ijF}*

Formal transit fares for a given route ij are calculated using the Cape Town MyCiti bus rapid transit distance-based fare scheme, where I again make the calculation using the straight-line distance between TAZ centroids.[38]

D.5.6 *Car Commute Cost*

I calculate the car monetary commute cost τ_A using data. First, I take an "average [monthly] 'total mobility cost' " calculated by the South African firm WesBank of ZAR 9,356.80; I divide this Figure by the number of working days per month (22.5), multiply by the share of all car trips made *to* work, 0.21, in the 2013 Cape Town survey travel diary, and convert to USD.[39]

D.5.7 *Minibus Operations Parameters δ_0, δ_1, χ*

I set the minibus shift length δ_0 and inverse "number trips" δ_1 to generate reasonable numbers of expected trips in half a typical work shift. I use half of a typical 8-hour work day as minibus drivers' time budget because my model depicts only a one-way commute: $\delta_0 = 240$ ensures that minibus drivers whose expected total trip time $\overline{\eta}/\iota_{ij} + 1/d_{ij}$ equals 240, i.e. consumes their entire work shift, complete one trip in expectation. In turn, I

[36] In each case, I use only the first suggested itinerary. The wait time is defined as the lag between the queried departure time and the time at which the first itinerary suggested by the API begins from the queried origin centroid. The specific date used is 25 Aug. 2021.

[37] Note that I set the one relation with zero estimated wait time to 1 min.

[38] The MyCiti scheme can be found here.

[39] Wesbank's figures can be found here.

assume that, even if total trip time equals zero, minibuses lose 20 minutes per trip for repositioning or driver breaks that are of course outside my model. As a result, under zero travel time, a minibus should be able to complete an average of 12 trips in the half-day shift that I model. Setting $\delta_1 = 0.0104$ exactly ensures that the expected number trips under $\overline{\eta}/\iota_{ij} + 1/d_{ij} = 0$ equals 12.

Next, I calibrate the minibus per-kilometer operating cost χ with figures provided by the firm GoMetro. I multiply the Toyota Quantum minibus's litres of diesel used per kilometer, 0.099, by the June 2022 diesel per-litre price in South Africa, ZAR22.63, convert to USD as with other prices in my model, and adjust by the ratio between my calibrated $\overline{\eta}$ and the actual bus capacity, 15, to set operating costs in proportion to revenue from the calibrated fares.

E. Results Appendix

E.1 Additional Validation

E.1.1 Mode Choice Probabilities

A standard yet relatively demanding test of the model is whether its predicted mode choice probabilities π^g_{ijm} by origin-destination pair and skill are correlated with those in the data. In Table E.1, I show that, for both minibus and car, the two are significantly positively correlated with nontrivial R-squared values. Note that the R^2 are not higher because, unlike standard spatial models, I do not back out any location-specific residuals from the data.

E.1.2 Mode Choice by Income

In the cross-section across origin-destination tuples, the model replicates the negative (positive) relationships between average minibus (car) mode shares and income quite well, as evidenced by Figure E.1.

E.1.3 Substitution Across Modes

I have shown that the model replicates the minibus matching process as well as the cross-sectional mode choice patterns and income gradients. It further reflects the fact that poor formal transit connections, as evidenced by higher formal transit travel times, lead

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

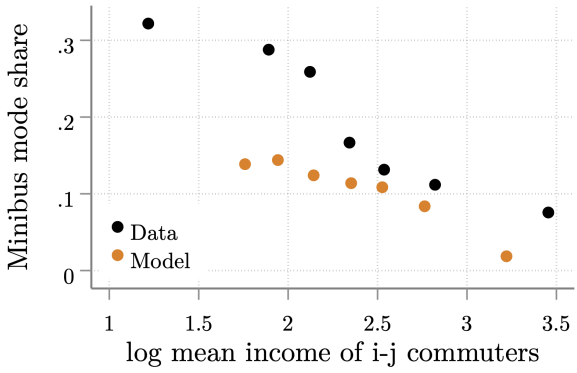
	Minibus	Car
Variables	Mode Share, Data	Mode Share, Data
Mode Share, Model	0.985***	1.025***
	(0.130)	(0.0804)
Constant	0.0580***	0.00425
	(0.0145)	(0.0497)
Observations	507	507
R-Squared	0.097	0.236

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

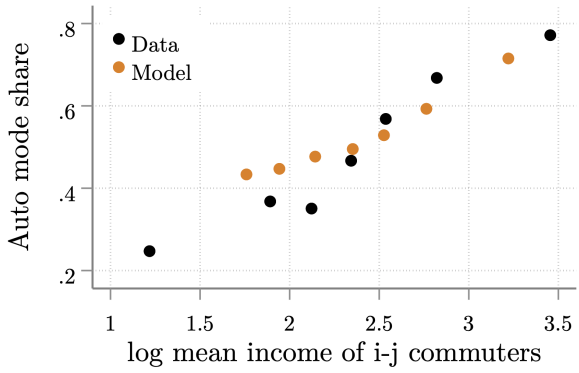
Notes: This table's unit of analysis is a skill group by origin by destination transport analysis zone tuple in Cape Town. In the data, I calculate skill-specific origin-destination transport analysis zone (TAZ) commute mode shares from the 2013 Cape Town Household Travel Survey, taking shares of respondents working outside the home who commute by each mode, where minibus includes all who use minibuses at some point during their commutes. The model probabilities are the route-skill-group-specific mode-choice probabilities π^g_{ijm}, which also correspond to pairs of TAZ.

FIGURE E.1. ORIGIN-DESTINATION MODE SHARES VERSUS INCOME

(A) MINIBUS MODE SHARE

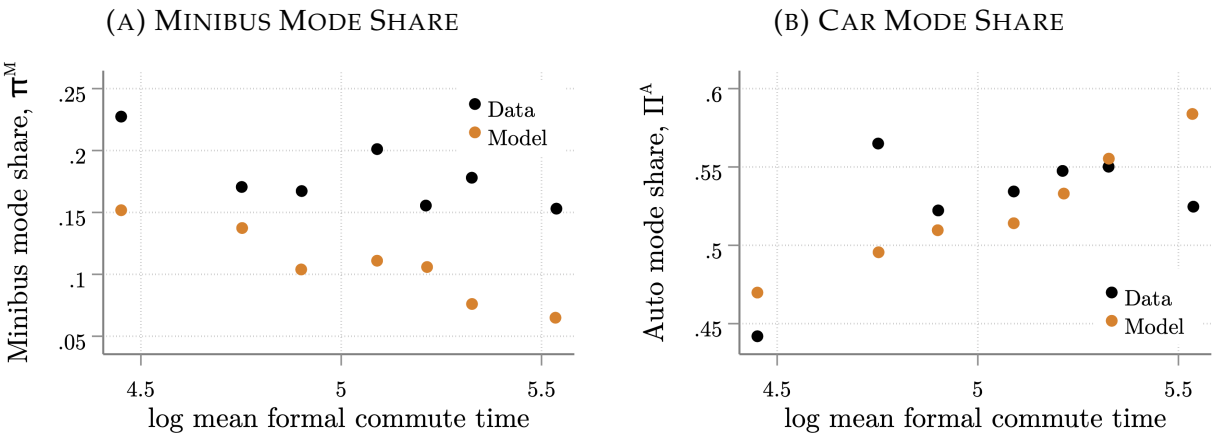


(B) CAR MODE SHARE



Notes: This figure's binned scatterplots display the relationship between mean income and mode shares, where the unit of analysis is an origin by destination transport analysis zone tuple in Cape Town. In the 2013 Cape Town Household Travel Survey data, for each origin-destination, I compute average household income per-person and (half) workday as well as the shares of respondents working outside the home who commute by each mode, where minibus includes all who use minibuses at some point during their commutes. In the model, I take a weighted average of the corresponding choice probabilities π^g_{ijm} and income ω^g_j over skill groups using inflows N^g_{ij}.

FIGURE E.2. ORIGIN-DESTINATION MODE SHARES VERSUS FORMAL COMMUTE (WAIT + TRAVEL) TIME



Notes: This figure's binned scatterplots display the relationship between the formal commute time and mode shares for commuters commuting between an origin and destination transport analysis zone tuple in Cape Town. Formal commute times, both in the data and in the calibrated model, are the sum of wait and travel times from the Azure API. In the 2013 Cape Town Household Travel Survey data, for each origin-destination, I compute the shares of respondents working outside the home who commute by each mode, where minibus includes all who use minibuses at some point during their commutes. In the model, I take a weighted average of the corresponding choice probabilities π^g_{ijm} and income ω^g_j over skill groups using inflows N^g_{ij}.

primarily to increases in driving, rather than minibus use (Figure E.2).[40] Thus, importantly, minibus demand appears to be quite inelastic.

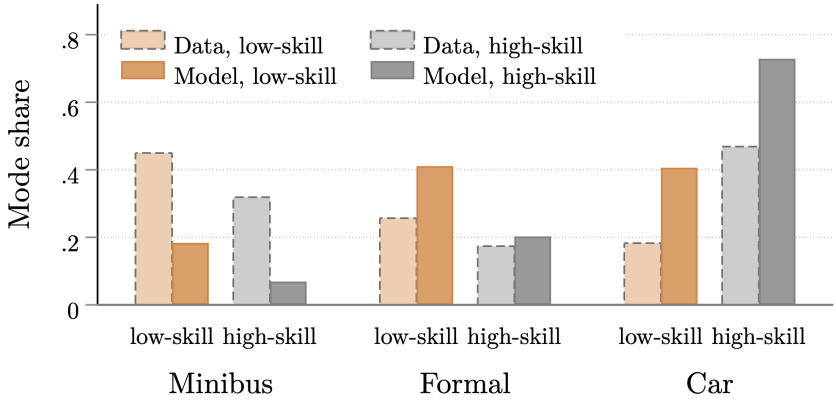
E.1.4 Stated Preference Approach

Stated preference data is only as useful as respondents' ability to predict their future non-hypothetical choices. I have already demonstrated in Section VI that the model matches aggregate mode shares and does so primarily due to the utility costs estimated from stated preferences; to further confirm the applicability of respondents' stated choices to their future decisions, I pursue two strategies.

First, I employ the model to predict the mode choices of the respondents in the combined stated preference sample. In particular, respondents' reported personal income and education, as well as the median model-calibrated wait and travel times for each mode, imply choice probabilities for each commute mode when inserted into the linearized (around $r = 0$) versions of equations (9)-(11). These predictions, of course, depend crucially on the estimated utility costs and other stated preference parameters. Figure E.3 demonstrates that my model, as estimated using respondents' choices, replicates the broad patterns

[40] Note that formal commute times are calculated using the Azure API, not from a commute survey.

FIGURE E.3. MODE SHARES, REPORTED VERSUS MODEL-PREDICTED IN COMBINED STATED PREFERENCE SAMPLE



Notes: This figure displays commute mode shares for the combined stated preference sample (from my own survey plus the 2013 Cape Town Household Travel Survey stated preference module), comparing respondent-reported (actual) commute modes with the mode shares predicted by the estimated model. In the data, minibus includes all who use minibuses at some point during their commutes. In the model, I take a weighted average of the corresponding predicted choice probabilities π^g_{ijm} across respondents.

in their actual mode choices. Thus, respondents were at least moderately successful in predicting their actual commute behavior, despite the potential difficulties associated with hypothetical choice questions. However, the model cannot fully explain the over-representation of minibus commuters in the stated preference sample noted in Appendix B.2.3.

Second, I re-estimate the discrete choice model and allow for heterogeneous effects by a series of demographic characteristics to explore the plausibility of any significant differences in preferences. I take equation (19) and interact a dummy variable h_i for the binary demographic characteristics listed in Column 1 of Table E.2 with the terms in the multinomial logit model that identify utility costs, their dependence on policy, and the value of time, as follows:

$$\pi^g_{icl} = \left(\sum_{l'} \exp\left(\overline{U}^g_{icl'}/\nu\right)\right)^{-1} \exp\Big[\zeta_{m(c,l)} + h_i\zeta_{m(c,l)} + \sum_z \left(\beta_z q_{cl}(z) + \beta^h_z h_i q_{cl}(z)\right) + \beta_{\text{time}}\omega_i\left(w_{cl}+t_{cl}\right) + \beta_{\text{time},h}h_i\omega_i\left(w_{cl}+t_{cl}\right) + \beta_{\text{fare}}\tau_{cl} + \beta_{\text{resid}}w_{cl}\tau_{cl}\Big] \quad \text{(E.1)}$$

The notation corresponds to that introduced previously, except that $\zeta_{m(c,l)}$ and β_z now indicate dummy variables for mode and the effects of minibus quality improvements z, respectively, that are not interacted with skill group except when h_i indicates college

TABLE E.2. STATED PREFERENCE HETEROGENEITY:
DIFFERENCE IN PARAMETER ESTIMATE, VERSUS BASE CATEGORY

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

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: This table displays parameters estimated from the multinomial logit model (E.1). Each cell gives the estimate, significance, and standard error of the effect of a dummy variable h_i that takes the value 1 for each demographic characteristic listed in the first column on the parameter listed at the top of each column. The unit of analysis is a mode alternative by choice set by individual respondent in either my newly collected minibus stated preference survey or a stated preference module of the 2013 Cape Town Household Travel Survey. The dependent variable is an indicator the respondent choosing a given choice set; the independent variables include indicators for minibus and for formal transit (MyCiti bus, bus, and train) interacted with the demographic dummy variable h_i, interactions of the minibus indicator with indicators for the presence of three quality improvements (again further interacted with skill) as well as h_i, monetary cost, cost interacted with wait time, and personal daily income interacted with the sum of wait and travel time as well as h_i. The mode-related attributes are defined in the hypothetical choice sets in the surveys; personal income is directly collected in my survey and imputed from the household section of the 2013 survey. Note that I normalize $\kappa_A = 0$ and restrict the sample to individuals employed outside the home between 25 and 65 years old.

workers. Women and college workers have a higher value of time saved, even conditional on income; the former result echoes Borghorst, Mulalic, and Ommeren (2021), who find that women's marginal cost of commuting increases after the birth of children. That college-educated workers should value their time more highly, even conditional on income, might similarly reflect a higher value of home production. Women have a *lower* utility cost of using minibuses or formal transit, or, equivalently, a higher utility cost of driving.

Surprisingly, women place a lower value on security; perhaps men are more likely to be targeted for reasons other than bodily strength or more likely to be intertwined in gang activities that would put them at risk.[41] Older workers, in turn, have a greater preference (lower utility cost) for formal transit, perhaps due to security risk. Indeed, they place a higher value on security and especially on driver adherence to speed limits, suggesting an

[41] Note that, in the final three columns, I display differences in the *absolute value* of the quality improvement effect.

intuitive greater risk aversion. Admittedly, the fact that older workers have a higher and women a lower preference for security is difficult to rationalize. On the whole, however, these results suggest that stated preference respondents did not randomly choose their answers simply to complete the survey. Instead, the plausibility of these choice patterns by demographic group suggests that they thought through their choices in a manner similar to that in which they might make actual commute mode choices.

E.2 Minibus Lanes: Implementation Details

I simulate minibus lanes on the top 10% of road links by minibus traffic, displayed in Figure E.4. Minibus traffic is defined as the inflows of buses on routes that use a given link:

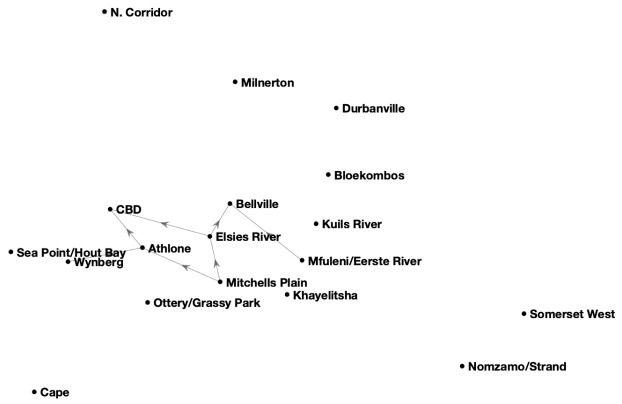
$$\begin{aligned}\text{minibus traffic}_{km} &= \sum_i \sum_j \text{bus inflow}_{ij} \mathbb{1}\{km \in \rho(i,j)\} \\ &= \sum_i \sum_j \left[d_{ij} \sum_g \left(\frac{p_{ij}^{T,g}}{\overline{\eta}} \right) \right] \mathbb{1}\{km \in \rho(i,j)\} \quad \text{(E.2)}\end{aligned}$$

To approximate the effects of taking these lanes away from cars, I continue to include minibuses in the calculation of the traffic volumes v_{ik} – which, on the targeted links, only affect cars' arriving-at-destination rates. To calculate the cost of minibus lanes, I sum the *Cost of Way* and *Land Cost - CBD/Commercial* categories, given in per-kilometer terms in De Beer and Venter (2021). I calculate the length of the shortest path between transport analysis zone (TAZ) centroids at 11pm using the Google Maps API with live traffic, and sum these lengths for all TAZ-TAZ links in the top 10% of links by traffic. I then multiply this total length by the cost per kilometer of adding a minibus lane in one direction, convert to USD, and multiply by the commuter rate of time preference to convert the one-time capital cost to a flow cost. This flow cost is then covered by equal lump-sum taxes by entering commuters.

E.3 Robustness

I now explore the robustness of my results, in particular on changes in parameters that reflect alternative institutional contexts. The latter include greater greater minibus association entry restrictions, higher red tape or association membership fees, and greater association market power. Furthermore, I consider an extension where minibuses can enter as informal to evade taxes.

FIGURE E.4. ROAD LINKS IN TOP 10% BY MINIBUS TRAFFIC



Notes: This map displays the top 10% of road network (transport analysis zone to adjacent transport analysis zone) links by model-predicted minibus traffic inflow.

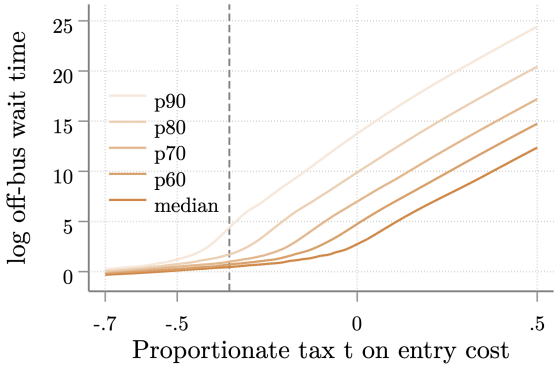
E.3.1 Entry Cost Elasticity ϕ

First, my estimated entry cost elasticity ϕ, as a form of summary statistic for a variety of forms of entry congestion, might not reflect the full extent of minibus associations' entry restrictions. To explore this possibility, I compute my results under an entry cost elasticity twice as high as my estimated value of $\phi = 0.0143$; in the limit, as $\phi \to \infty$, the model would approach the zero-entry case. Table E.3 summarizes all robustness checks. In this case, the optimal subsidy remains similar, but the higher entry cost congestion actually lowers entry costs on the low-demand, long routes, with $b_{ij} < 1$, that drive my results; as a result, initial off-bus wait times at the upper end of the distribution are disproportionately lower (Figure E.5a) so that commuters gain less from additional bus entry on these routes (Figure E.5b). Despite this mathematical effect, the overall picture remains similar.

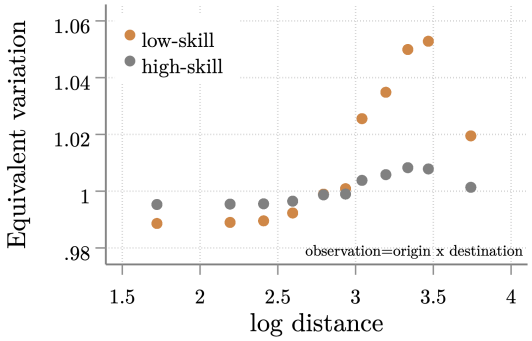
Matching efficiency, on the other hand, turns out to be more beneficial under restrictive entry because, though all routes see similar proportional off-bus wait time reductions as under my estimated ϕ (Figure E.6a), the lower initial wait times previously discussed mean that commuters benefit more from these proportionate wait time reductions due to the intuitive concavity of utility in wait time introduced by the extreme value shocks (Figure E.6b).

FIGURE E.5. ENTRY SUBSIDY ROBUSTNESS: UNDER 100% HIGHER ϕ

(A) OFF-BUS WAIT TIME PERCENTILES



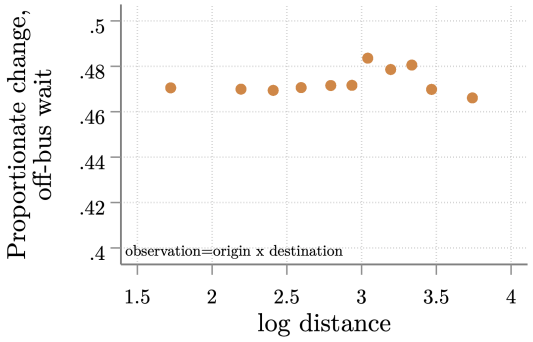
(B) EQUIVALENT VARIATION BY ORIGIN-DESTINATION AND SKILL VS. DISTANCE



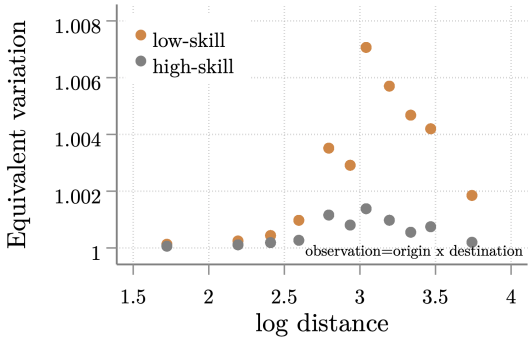
Notes: This figure reflects the model with 100% higher ϕ. Panel (A) displays the percentiles, across routes, of expected off-bus wait time of minibus passengers, calculated according to (4) at different levels of a proportionate tax t on minibus entry costs, which rise to $(1+t)\,\overline{\psi}b_{ij}^{\phi}$. Panel (B) displays the equivalent variation from the optimal subsidy $t = -0.32$, by skill as well as origin-destination, versus distance from i to j.

FIGURE E.6. MATCHING EFFICIENCY ROBUSTNESS: UNDER 100% HIGHER ϕ

(A) PROPORTIONATE CHANGE, OFF-BUS WAIT TIME VS. DISTANCE



(B) EQUIVALENT VARIATION BY ORIGIN-DESTINATION AND SKILL VS. DISTANCE



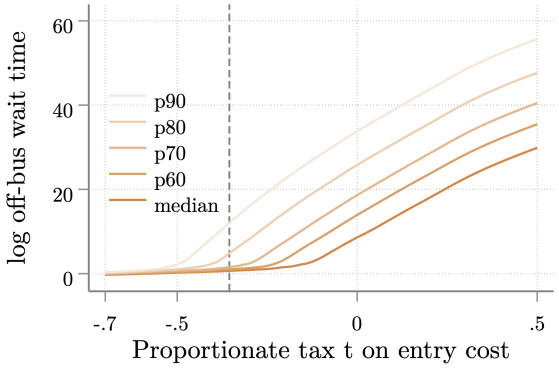
Notes: This figure reflects the model with 100% higher ϕ. Panel (A) displays the proportionate change in expected off-bus wait, $1/\lambda_{ij}$, from increasing matching efficiency, relative to baseline, against straight-line distance from i to j; Panel (B) displays the equivalent variation from increasing matching efficiency, by skill as well as origin-destination, versus distance from i to j.

E.3.2 Entry Cost "Intercept" $\overline{\psi}$

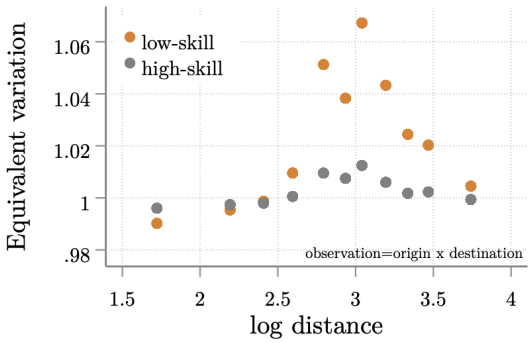
Second, imagine that bureaucratic red tape or the entry fees charged by minibus associations actually exceed my internally calibrated entry cost intercept $\overline{\psi}$. I increase the

FIGURE E.7. ENTRY SUBSIDY ROBUSTNESS: UNDER 20% HIGHER $\overline{\psi}$

(A) OFF-BUS WAIT TIME PERCENTILES



(B) EQUIVALENT VARIATION BY ORIGIN-DESTINATION AND SKILL VS. DISTANCE



Notes: This figure reflects the model with 20% higher $\overline{\psi}$. Panel (A) displays the percentiles, across routes, of expected off-bus wait time of minibus passengers, calculated according to (4) at different levels of a proportionate tax t on minibus entry costs, which rise to $(1+t)\,\overline{\psi} b_{ij}^{\phi}$. Panel (B) displays the equivalent variation from the optimal subsidy $t = -0.32$, by skill as well as origin-destination, versus distance from i to j.

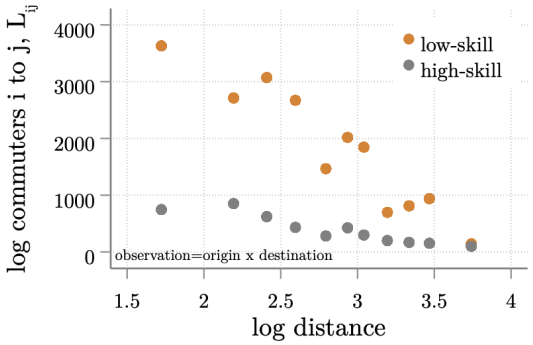
estimated $\overline{\psi} = 33.5$ by 20%. The optimal entry subsidy also changes little, relative to that under the estimated parameters, but the gains are significantly larger. Initial off-bus wait times rise across the distribution (Figure E.7a) due to lower bus entry; the highest gains from the entry subsidy then accrue to *shorter* routes than in the estimated model (Figure E.7b) because the routes that previously gained now have wait times so long that these commuters' utility is virtually flat in wait time. The routes that now gain, in turn serve larger numbers of commuters, since the exogenously calibrated commute flows logically decrease with distance, as in Figure E.8. As a result, aggregate equivalent variation (Table E.3) is higher for both skill groups.

The gains from improved matching rise by a similar amount as in the ϕ robustness exercise, but for a different reason that mirrors the preceding discussion. Once again, all routes experience similar proportional wait time reductions, but just as in the entry subsidy case, higher initial off-bus waits mean that the previously-marginal routes that benefit the most (Figure E.9b) now serve more commuters.

E.3.3 Fare Demand Slope, Γ_2

Third, my estimates of the reduced-form supply curve might understate the extent of minibus association market power. I explore the case of higher market power by increasing

FIGURE E.8. TOTAL COMMUTE FLOWS N_{ij}^{g} VERSUS DISTANCE

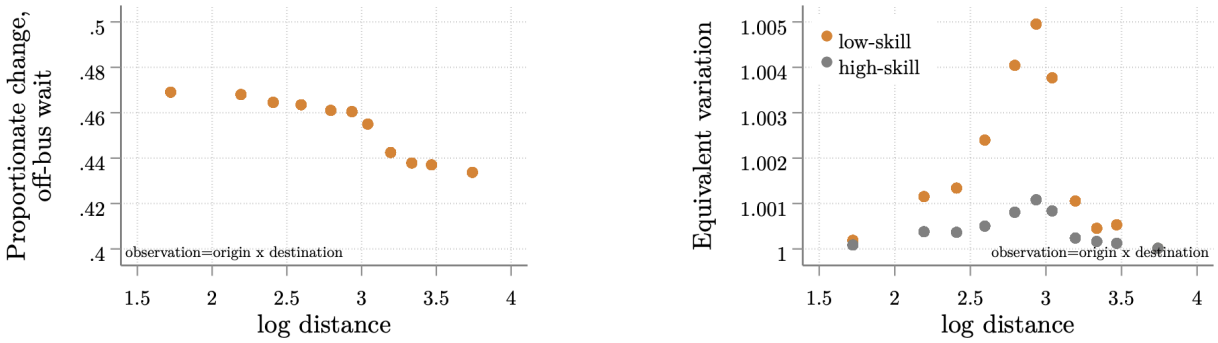


Notes: This figure displays the total (exogenously calibrated) commute flow N_{ij}^{g} by skill as well as origin-destination versus straight-line distance from i to j.

FIGURE E.9. MATCHING EFFICIENCY ROBUSTNESS: UNDER 20% HIGHER $\overline{\psi}$

(A) PROPORTIONATE CHANGE, OFF-BUS WAIT TIME VS. DISTANCE

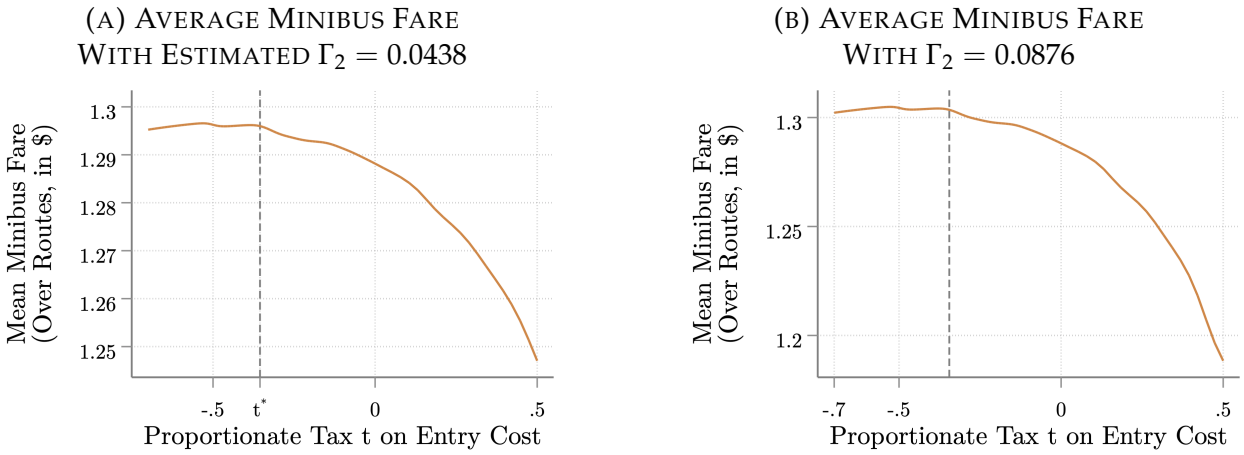
(B) EQUIVALENT VARIATION BY ORIGIN-DESTINATION AND SKILL VS. DISTANCE



Notes: This figure reflects the model with 20% higher $\overline{\psi}$. Panel (A) displays proportionate change in expected off-bus wait, $1/\lambda_{ij}$, from increasing matching efficiency, relative to baseline, against straight-line distance from i to j; Panel (B) displays the equivalent variation from increasing matching efficiency, by skill as well as origin-destination, versus distance from i to j.

the slope Γ_2 of supply, i.e. the sensitivity of fares to city-wide demand, by 100% relative to the estimated $\Gamma_2 = 0.0438$. I then adjust the fare intercept Γ_0 so that baseline equilibrium fares remain unchanged. The small shifts in aggregate minibus demand (recall Figure A.6b), however, mean that even a steeper fare curve, i.e. greater minibus market power, does not translate into a much larger counterfactual fare increase, as Figure E.10b demonstrates. The optimal subsidy in Table E.3 thus falls only slightly and the gains remain unchanged.

FIGURE E.10. FARES UNDER ENTRY SUBSIDY/TAX, AS ESTIMATED AND UNDER 100% HIGHER FARE DEMAND SLOPE Γ_2



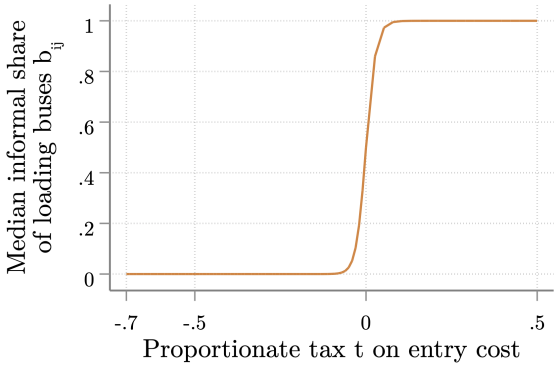
Notes: This figure displays the average minibus fare τ_{ijM} across routes at different levels of a proportionate tax t on minibus entry costs. Panel (A) does so at the estimated value of the fare sensitivity to demand Γ_2, while in Panel (B), I assume a value of Γ_2 twice as high.

E.3.4 Extension: Minibus Informality

City-run studies suggest that around half of minibuses in Cape Town lack official government permits (*Operating Licence Strategy 2013-2018* 2014, 77). My optimal entry tax or subsidy exercise, of course, presupposes the government's ability to successfully tax or subsidize *all* buses. To account for a potential lack of such state capacity, I solve for the optimal entry tax or subsidy in an altered model with two minibus sectors that operate on each route, namely legal and informal. I now detail the ways in which this alternative "informality" model differs from that in the main text. From the passenger perspective, both types of minibuses are equivalent; the quantity relevant for matching is the total mass of seaching buses $b_{ij} \equiv b^l_{ij} + b^i_{ij}$ where the superscripts l and i correspond to legal and informal minibuses, respectively. Similarly, road congestion is a function of the total minibus inflow. Since virtually all minibuses, both legal and informal (Antrobus and Kerr 2019), belong to the same associations, I impose that both types of minibus charge the same fares, according to (13).

The two sectors, legal and informal differ only in their entry costs. To generate an equilibrium with both legal and informal minibuses, I assume that the entry cost "intercept" differs by sector and that the congestion in the entry cost depends on the mass of buses in that sector only. Such an assumption is reasonable if this congestion reflects, at least in part, bureaucratic impediments to additional entry. Crucially, only legal minibuses pay

FIGURE E.11. MEDIAN INFORMAL SHARE OF LOADING MINIBUSES



Notes: This figure displays the median informal share of loading minibuses, $b_{ij}^i / \left(b_{ij}^i + b_{ij}^l\right)$ across routes at different levels of a proportionate tax t on *legal* minibus entry costs, in my informality extension.

any tax or receive any subsidy t. Sector s entry costs thus equal $(1 + \mathbb{1}\{s = l\}\, t)\, \overline{\psi}_s \left(b_{ij}^s\right)^{\phi}$. Of the equations that determine equilibrium, only the free entry conditions change; in addition to $b_{ij} \equiv b_{ij}^l + b_{ij}^i$, the sector-specific bus loading masses b_{ij}^s must satisfy

$$\left(\overline{\eta}\tau_{ijM} - \chi\Delta_{ij}\right) \exp\left[-\delta_1\left(\frac{\overline{\eta}}{\iota_{ij}} + \frac{1}{d_{ij}} - \delta_0\right)\right] = (1 + \mathbb{1}\{s = l\}\, t)\, \overline{\psi}_s \left(b_{ij}^s\right)^{\phi}. \tag{E.3}$$

To match equilibrium median relative bus entry, $\frac{b_{ij}}{p_{ij}}$, as in the main calibration, and the informal share of minibuses in the data, I set $\overline{\psi}_s = 1.01\overline{\psi}$, where $\overline{\psi}$ is the calibrated entry cost from the main model.

Were the standard model to imply an optimal tax $t > 0$, this informality model would produce vastly different results. However, given that my proposed policy in fact benefits minibus entrants, the optimal tax and associated welfare gains are unaffected; imposing a subsidy $t < 0$ rapidly sends the median informal share of minibuses across routes to zero, as in Figure E.11.

TABLE E.3. POLICY COUNTERFACTUALS UNDER DIFFERENT PARAMETER ASSUMPTIONS

	Entry Subsidy			Matching Efficiency		Security		Lanes	
		EV by Skill		*EV by Skill*†		*EV by Skill*		*EV by Skill*	
Parameter Change	t^*	Low	High	Low	High	Low	High	Low	High
As Estimated	-0.36	1.007	0.999	1.001	1.000	1.025	1.011	1.009	1.000
Increase ϕ *by 100%*	-0.32	1.006	0.999	1.002	1.001	1.025	1.011	1.009	1.000
Increase $\overline{\psi}$ *by 20%*	-0.29	1.02	1.003	1.002	1.001	1.022	1.01	1.016	1.001
Increase Γ_2 *by 100%*	-0.34	1.007	0.999	1.001	1.000	1.026	1.011	1.010	1.000
Informality Extension	-0.36	1.007	0.999						

†: does not account for cost of policy, unlike other counterfactuals.

Notes: This table displays, for alternative parameter or modeling assumptions, implemented one at a time: the optimal entry subsidy t^*, the (after-tax) equivalent variation from implementing this optimal subsidy, the pre-tax equivalent variation from the matching efficiency increase counterfactual, as well as the post-tax equivalent variation from the security guard and exclusive minibus lane counterfactuals, all relative to the equilibrium without the corresponding policy. "As estimated" indicates the parameters under which I solve the model and implement counterfactuals in the main text, displayed in Table 1.