Fare Structure and the Demand for Public Transit

Yizhen Gu Qu Tang Yacan Wang Ben Zou*

September 2023

Click HERE for the most recent draft

Abstract

Public transit prices are often complex and are designed to reflect multiple policy goals. This paper evaluates transit users' responses to a comprehensive fare adjustment in the Beijing Subway that replaced a flat rate with one that varies by distance, time, and month-to-date cumulative expenditure. We find the demand for subway trips is inelastic, travel schedules are inflexible, and users are largely unresponsive to cumulative quantity discounts. We consider several alternative revenue-equivalent fare structures and show that their aggregate and distributional welfare impacts depend crucially on how users perceive and react to the complex fare structure.

Keywords: Public transit, demand for travel, price elasticity

JEL Codes: R41, R48, L92, L98, D12

*Gu: HSBC Business School, Peking University; email: yizhengu@phbs.pku.edu.cn. Tang: Institute for Economic and Social Research, Jinan University; email: qutang@jnu.edu.cn. Wang: Beijing Jiaotong University; email: ycwang@bjtu.edu.cn. Zou: Purdue University; email: zou136@purdue.edu. We thank Liming Chen, Victor Couture, Gilles Duranton, Lindsay Relihan, and seminar and conference participants at Purdue University, University of Maryland, Jinan University, Zhejiang University, Peking University, Chinese University of Hong Kong, Shenzhen, Fudan University, Shanghai University of Finance and Economics, Mid-Midwest Applied Conference, Northeast Environment and Energy Economics Conference, SMU Conference on Urban and Regional Economics, and ADB Conference on Infrastructure and Urban Development for helpful comments. We are grateful to the people at the Beijing Institute of City Planning for providing and helping us with the data. We have benefited from outstanding research assistance by Yiyao Zhou. Gu gratefully acknowledges funding from the Peking University Shenzhen Graduate School (Grant No. 1270110213) and the Natural Science Foundation of China (Grant No. 72273008). All views expressed and errors are our own.

1 Introduction

Public transit is indispensable in many cities around the world. In particular, policymakers in fast-growing developing country megacities hope that rapid, reliable public transit can alleviate congestion, provide access to jobs and amenities, and offer a viable alternative to a car-based urban structure (World Bank, 2009). To make the transit system work effectively, it is important to get the price right.

Often heavily regulated, transit operators aim to achieve multiple goals, including the system's aggregate efficiency, distributional impacts, and potential externalities on surface roads. They also face many practical constraints, such as a limited budget. Multiple incentives along different dimensions are built in, resulting in complex fare structures (Cervero, 1990). However, how transit users, who must make on-the-spot decisions in a typically hectic environment, respond to a complex fare structure remains an open question.

Existing theoretical and empirical work on transit pricing provides limited guidance. Existing theories have focused on overall efficiency under a single-dimension fare level and perfectly rational consumers (Small and Verhoef, 2007; Parry and Small, 2009), while empirical studies typically use aggregate data that inevitably misses rich heterogeneity (Davis, 2021). The lack of consensus on the optimal fare structure is also reflected in the distinctly different fare structures adopted by major transit systems across the world, from London Tube's labyrinthine schedule with multi-dimensional incentives to New York Subway's single flat rate.

This paper provides the first comprehensive study on how fare structures affect user behavior and determines the transit system's welfare, efficiency, and externality. The empirical setting is the Beijing Subway, one of the world's largest urban rail transit systems. We study a complete overhaul of its fare structure that replaced a flat rate with one that consists of three major components: (1) single-trip fare is a step function of distance; (2) trips that start before the morning peak hour are qualified for an "early-bird discount" (EBD); (3) frequent users qualify for discounts, where the discount rate is a step function of the month-to-date expenditure on the subway. Our main data are the universe of trip-level records captured by an electronic payment system in periods both before and after the fare adjustment.

The fare adjustment provides useful variations along several dimensions, which allow us to estimate key parameters capturing user responses to the new fare structure. The *demand elasticity* measures the extent to which users take fewer trips as the price increases. The *rescheduling elasticity* measures the degree to which users are willing to reschedule trips to qualify for a lower price. We test how consumers respond to kinks in the budget constraint generated by the cumulative quantity discounts. Those are key to describing transit users' behavior and have generalizable interpretations.

We find both demand and rescheduling elasticities small and limited evidence that users respond to the discounts. While the new fare structure is effective in raising revenue, it causes large utility losses to the users, especially those who use subway frequently. The aggregate and distributional welfare impacts also depend crucially on how users perceive and respond to the non-linear pricing, as does the design of the optimal fare structure. These findings are uncovered in several steps.

The new stepped pricing of single trips generates sharp discontinuities in fare between origin-destination (OD) station pairs that differ only slightly in distance and motivates a regression discontinuity (RD) design . Precisely because of the sharp difference in fare, passengers may strategically shorten their trips. We design several approaches to account for such strategic bunching. Alternative approaches generate similar results. The bunching-adjusted demand elasticity is around -0.36.

While the data do not have transit riders' demographics, users can be classified by their ridership patterns. The largest demand elasticity is found among regular commuters. Within each user type, the elasticity is similar across trips during different times and possibly for different purposes. This suggests that users do not systematically substitute between different kinds of trips.[1]

The progressive discount rate based on cumulative expenditure creates a non-trivial optimization problem for the user. We distinguish four behavioral types in response to the non-linear budget constraint. A consumer is "rational" if she fully foresees her travel demand and optimizes based on the end-of-month marginal price. Instead, "ironers" respond to the average price.[2] A consumer is "myopic" if she makes decisions based on the instantaneous marginal price she currently qualifies for. Finally, we call a consumer "oblivious" if she only responds to the listing price.

Under the assumption of a smooth distribution of preferences, there will be a "bowl" in the distribution of monthly subway expenditures around a non-convex kink in the budget (Saez, 2010; Kleven, 2016). We find no evidence for such non-smoothness in the distribution and reject users being fully or approximately rational. Myopia is also ruled out via a simple, intuitive test. We then estimate the composition of behavioral types by building the statistical mixture model in our station-pair RD framework. We consistently find that users respond only to the listing price. This is true even among frequent users who have a regular travel pattern.

The rescheduling elasticity is estimated by exploiting the EBD and a difference-in-discontinuity design. There is a clear bunching of trips right before the time cutoff (which qual-

[1]For example, in response to the higher fare, commuters may cut weekend travels but save commutes, in which case we would expect a larger demand elasticity for weekend trips than for commutes.

[2]They are called ironers because they treat the kinked budget constraint as if it is ironed smooth and linear. See Liebman and Zeckhauser (2004) for the original use of the term.

ifies for a lower fare) and a trough right after. However, the trough is neither deep nor wide. The rescheduling elasticity peaks at around -0.4 for trips immediately after the time cutoff but quickly reduces to zero for trips that are 15 minutes after. Overall, a 30% EBD diverts less than 1% of peak-hour trips to a different time.

Empirical results are used to evaluate the welfare impacts of the fare adjustment. The listed fare for an average trip increased from 2 yuan to 4.7 yuan. With users not responding to cumulative quantity discounts, revenue increases by 56% at the cost of a 25% decline in ridership. Assuming zero marginal operating cost, the deadweight loss (the sum of the changes in revenue and consumer welfare) amounts to 63% of the increased revenue. Welfare loss disproportionately falls on regular commuters, who cut trips substantially and forgo large discounts by failing to optimize on the kinked budget. Reduced subway ridership also generates negative externality on surface roads, resulting in substantial congestion externalities equivalent to 41% of the increased revenue.

Welfare impacts would differ substantially had users responded differently to the cumulative quantity discounts. With fully rational users, the ridership would be 15% higher and the deadweight loss 58% smaller than that with completely oblivious users. Welfare implications with heuristic optimizers lie between the two polar cases.

Finally, we simulate welfare impacts under alternative pricing schemes. While in theory, the current fare structure generates a good balance in efficiency and distributional goals, with users less than fully rational, a simple revenue-preserving flat rate results in higher social welfare. On the other hand, a rush-hour surcharge substantially reduces commuting trips but diverts few to less crowded hours. As a result, it generates high costs due to increased congestion externality and consumer welfare losses, especially among regular commuters.

Related Literature

A well-functioning public transit is essential to make the city work. Recent studies show that transits alleviate road congestion (e.g., Anderson, 2014; Gu et al., 2021b) and air pollution (e.g., Chen and Whalley, 2012; Li et al., 2019; Gendron-Carrier et al., 2022), improve access to jobs and amenities (e.g., Lu et al., 2021; Zárate, 2022), and shape urban geographic structure in the long run (Heblich et al., 2020; Tsivanidis, 2019; Balboni et al., 2020).

However, there are few studies on the pricing and demand for public transit.[3] Existing theories focus on aggregate efficiency with an optimal single-dimensional fare level (Vickrey, 1980; Small and Verhoef, 2007; Parry and Small, 2009), abstracting away from important practical considerations such as budget constraint, distributional impacts, and behavioral responses. Previous

[3]This is particularly surprising given there is a large literature on road pricing since Vickrey (1963)

empirical work mostly relies on aggregate ridership data (e.g., Davis, 2021), and is unable to unpack rich heterogeneity in consumers' responses and welfare consequences.[4] We contribute to this literature by conducting comprehensive evaluations of a realistic fare structure by exploiting a unique policy experiment, using the universe of trip-level records, and adopting tailored empirical designs to credibly identify key economic parameters.

The findings in this paper have broad implications for how consumers respond to complex pricing schedules. The pricing of many regulated natural monopolies, such as electricity and water, share similarities with public transit. The price structures of such goods are often designed to smooth demand over time (Jessoe and Rapson, 2014), account for externalities (Reiss and White, 2005), and consider distributional impacts (Borenstein, 2012). Moreover, there has been growing literature on behavioral responses to complex pricing schemes. Previous studies have found that consumers do not rationally respond to the marginal price (e.g., Borenstein, 2009; Ito, 2014); they would benchmark their responses on past choices (Ito and Zhang, 2020), and are often unaware of or unable to fully absorb complete information (Sexton, 2015). This paper finds that behavioral responses are important considerations in designing transit fare structure and have consequential efficiency and welfare implications.

The rest of the paper is organized as follows. Section 2 describes Beijing's subway system, the background of the fare adjustment, and the data used in this paper. Section 3 presents empirical strategies and findings. Section 4 evaluates the aggregate and distributional impacts of the fare structure change. Section 5 discusses whether alternative fare structures could achieve better outcomes. Section 6 concludes.

2 Background and Data

2.1 The Beijing Subway

Beijing is China's capital and its second-largest city by population. Its sprawling metro area has a population of 24 million and extends into neighboring provinces (Chen et al., 2022). Rapid population growth, urban expansion, and rising car ownership in the past few decades have made Beijing one of the world's most congested cities. In 2015, the average one-way commute was 12 kilometers long and took about 60 minutes (Gu et al., 2021a).

To alleviate congestion and provide access, the Beijing government has invested heavily in the city's subway network in the past two decades. In 2001, the system had two lines and about

[4] As many transit systems adopted automatic fare collection methods, recent empirical studies have taken advantage of trip-level records. Most of those studies are in the transportation literature (e.g., Pelletier et al., 2011; Ma et al., 2020). Those studies typically do not have an empirical design for causal identification and do not provide normative implications of the fare structure design.

40 stations. By 2019, the system had 25 lines, over 460 stations, and an annual ridership of 3.8 billion. Excluding walking trips, the subway accounted for 15% of the commuting trips and about 40% of the passenger mileage in 2014 (Beijing Transport Institute, 2015). A typical subway trip was 15 kilometers (km) long and took about 37 minutes, including waiting time (Appendix Table A.1). The average speed of a subway trip is about 24 km per hour, comparable to private car trips and much faster than bus or bike trips.

Government-owned companies operate the Beijing Subway.[5] Transit fare is determined by the municipal government and is heavily subsidized. Before the fare adjustment in 2014, Beijing's subway had a flat-rate fare of 2 yuan (approximately 0.3 US dollars) for one trip regardless of distance and time of travel. This rate was first adopted in 1996. The subway, the city's population, and its economy had all substantially grown during this period. With ridership and operational costs rising, the subway system was losing several billion yuan yearly.

2.2 Subway Fare Adjustment

In December 2014, the Beijing Subway overhauled its fare structure. The simple flat rate was replaced by a more complex pricing schedule with three major components. First, the listing price for a single trip is a step function of track distance. It starts from 3 yuan for a ride under 6 km, and incrementally rises to 10 yuan for a ride between 92 and 112 km. Second, users qualify for a cumulative-quantity discount if they use the electronic smartcard for payment. A user starts the calendar month by paying the listing price but starts to qualify for a 20% discount for her next trips once she has spent 100 yuan on the subway. The discount rate rises to 50% once the out-of-pocket expenditure reaches 150 yuan. The discounts are capped. Once the out-of-pocket expenditure reaches 400 yuan, the user receives no further discount for the remainder of the month. Third, starting in December 2015, An early-bird discount (EBD) was applied in 16 often crowded stations during morning peak hours. A 30% discount is applied to the fare if a passenger enters one of those stations before 7 AM.

Table 1 summarizes the fare adjustment. Had the ridership structure remained unchanged from that in September 2014, the post-reform average fare for a single trip would have been 4.6 yuan at the listing price and 4.3 yuan after the discount. It was a substantial rise. Consider a regular user who rides the subway twice every weekday. Her monthly expenditure on subway

[5]Beijing's public transit also includes an extensive bus system consisting of 1,158 bus lines and a fleet of 25,000 buses. Bus ridership has been declining in recent years due to rising car ownership and the rapid expansion of the subway. Nevertheless, there were still ten million daily bus trips by 2019 (Beijing Transport Institute, 2015). Buses are slow, and bus trips are typically short. The median trip was just under 2 km and took about 20 minutes (Beijing Transport Institute, 2015). Bus trips cost much less than subway trips of the same length. In 2014, most bus trips cost 0.5 yuan for smartcard users. Bus fares were doubled on the same day of the subway fare adjustment. But the bus remains a far more affordable option than the subway. This paper focuses on the impact of subway fare adjustment on demand for subway trips and does not consider interactions between the subway and the bus.

trips will be about 150 yuan or 4% of the monthly earnings of an average full-time worker in Beijing in 2015. It was also a comprehensive fare structure change that incorporated price variation in multiple dimensions, which allows for separate identification of key parameters that govern consumer demand.

The new fare structure was motivated by several policy goals.[6] The first was to raise revenue while keeping a high level of ridership. Whether the fare rise can achieve this goal hinges on demand being inelastic. The second goal was to smooth the temporary distribution of the trips and reduce crowdedness during peak hours. The EBD was intended to achieve this goal by creating a price difference over time. How much it is successful depends crucially on the flexibility of users' travel schedules, captured by the rescheduling elasticity. The third goal was relatively fair distributional impacts. While frequent subway users and those with long trips stand to pay a lot more, the cumulative quantity discounts aim at alleviating their financial burden. Those discounts also create kinks in the monthly budget. Whether that goal can be achieved depends on whether users can correctly perceive and rationally respond to the nonlinear budget.

While transit systems across the world face different constraints and adopt different fare structures in practice, the impacts on revenue, ridership and its temporal pattern, distributional welfare, and externality on surface roads are common considerations for the design of the transit fare Cervero (1990). While in theory, balancing these goals mandates a mixture of multiple pricing tools, some transit authorities are wary of suboptimal behavioral responses induced by a complex fare structure and opt for a simple, straightforward pricing schedule.

2.3 Data

Beijing's public transit uses an electronic smartcard as the storage and payment method. It is widely popular because it is easy to use and because card users are qualified for discounts.[7] By comparing the official statistics with our data, we estimate that around the time of the fare adjustment, between 90% to 95% of the subway trips were paid for by a smartcard. The smartcard captures the entry and exit times and locations for each trip. Trips can be linked via an anonymized card number.

The main data used in this paper are the universe of subway trips that used a smartcard and took place in the week between September 15 and September 21, 2014 (before the fare adjustment), the full month of April 2015, and the week between September 12 and September 18, 2016 (used to evaluate the impacts of the EBD). During this period, the average number of daily trips in our

[6] See, for example, the following news article (in Chinese) for an interview with a city government official in the department of Public Transportation. http://politics.people.com.cn/n/2013/1219/c1001-23885064.html (last accessed on September 2, 2023).

[7] The smartcard can also be used on buses, where the user automatically qualifies for a 50% discount on all trips.

Table 1: Summary of Subway Fare Structures

Before Dec. 28, 2014				
flat rate of 2 yuan for all trips				
After Dec. 28, 2014				
	distance-based	cumulative quantity discounts		
		expenditure-to-date (yuan)		marginal
distance (km)	fare (yuan)	before discount	after discount	discount rate
$[0, 6]$	3	$[0, 100]$	$[0, 100]$	0
$(6, 12]$	4	$(100, 162.5]$	$(100, 150]$	30%
$(12, 22]$	5	$(162.5, 662.5]$	$(150, 400]$	50%
$(22, 32]$	6	> 662.5	> 400	0
$(32, 52]$	7			
$(52, 72]$	8			
$(72, 92]$	9			
$(92, 112]$	10			
After Dec. 26, 2015				
	early-bird discount			
condition		discount rate		
entering one of the 16 stations*		30%		

* The 16 stations are on the Batong Line and the Changping Line.

data is about 4.5 million on weekdays and 3.2 million on weekends and holidays. Appendix Table A.1 provides more detailed summary statistics.

We geocode subway stations and collect the track distance and fare between each station pair from Beijing Subway's website. There are 354 stations in our sample. About 67,000 origin-destination station pairs have non-zero ridership in the September 2014 data; there are about 112,000 such pairs in the April 2015 data.

3 Estimating Responses to the Fare Adjustment

This section explores consumers' responses to the fare structure adjustment along several dimensions. We first estimate the demand elasticity for subway trips by exploiting price discontinuities in distance. While trip-level data do not have demographic information, we classify users by travel patterns and estimate heterogeneity in demand elasticity by user type. We then investigate behavioral responses to the cumulative quantity discounts. Finally, we estimate the rescheduling elasticity by exploiting the price discontinuity in time created by the EBD.

3.1 Demand Elasticity

3.1.1 Stepped Pricing and the Regression Discontinuity Design

The step-wise increases in fare by distance provide a natural setting for a regression discontinuity (RD) design. Figure 1 illustrates the empirical design with the first four distance cutoffs. Each dot represents station pairs that fall within a 500-meter distance bin. The $y-$axis shows the *log change* in ridership between 2014 and 2015. Red vertical lines indicate distance cutoffs. There are abrupt declines for trips in distance bins that are immediate to the right of the cutoffs.

The size of the decline at the cutoff indicates the magnitude of demand elasticity. We estimate the following equation around each distance cutoff:

$$\Delta \ln(N_{od}) = e^{RD} \cdot \Delta \ln p_{od} + \boldsymbol{f}(dist_{od}) + \Phi_{o,d} + \Delta \varepsilon_{od} \tag{1}$$

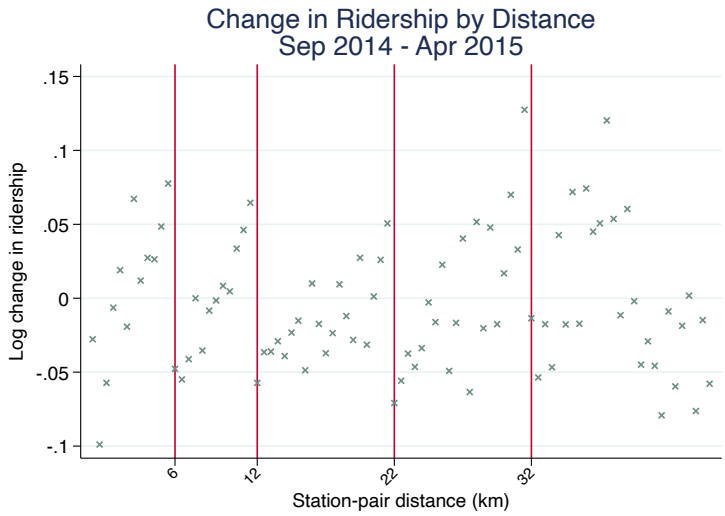
Each observation is an origin-destination (OD) station pair with non-zero ridership in September 2014. OD pairs are assigned to each cutoff in non-overlapping windows. The dependent variable is the log change in ridership between 2014 and 2015. $\boldsymbol{f}(dist_{od})$ represents flexible functions of the running variable. In the baseline, we use linear functions of distance separately for either side of the cutoff. $\Delta \ln p_{od}$ is the log change in the listing price of the OD pair. It is equivalent to including just the log new listing price because the initial price equals 2 for all OD pairs. $\Phi_{o,d}$ is a vector of control variables that account for the heterogeneity in ridership changes and make the estimate more precise. In the baseline, it includes the origin and the destination fixed effects. The coefficient associated with $\Delta \ln p_{od}$ can be interpreted as the demand elasticity. Observations are weighted by each OD pair's ridership in September 2014, so the estimated demand elasticity applies to the aggregate ridership.

The results of estimating Equation 1 at the first four distance cutoffs are reported in Columns 1 through 4 of Table 2. The estimates of the elasticity are all highly statistically significant and are tightly distributed between -0.35 and -0.45. The magnitude of the elasticity does not appear to be a monotone function of distance.[8] Therefore, for the rest of the paper, we pool all cutoffs together by estimating Equation 1 with all OD pairs included. For the baseline, $\boldsymbol{f}(.)$ is a 5^{th}-order global polynomial of distance. Column 5 shows that the overall demand elasticity is -0.39.

We conduct common checks to test the validity of the RD design (Cattaneo et al., 2019), as well as a host of robustness checks, which include varying the degree of the distance polynomial and using the local polynomial RD estimation with optimal bandwidth (Calonico et al., 2014). The baseline demand elasticity is remarkably robust to those checks. Appendix B.1 provides details of the validity tests and robustness checks.

[8] For example, one may expect the elasticity to be smaller for longer trips because there are fewer good substitutes for the subway.

Figure 1: Change in Station-pair Ridership



Note: Each dot represents the log change in ridership in all OD pairs within a 500-meter distance bin between September 2014 (populated to the full month) and April 2015. Red vertical lines indicate distance cutoffs for a higher listing fare. See Appendix Figure B.1 for the ridership by distance bins separately in September 2014 and April 2015.

Table 2: Demand Elasticity for Subway Trips

	(1)	(2)	(3)	(4)	(5)	(6)
					all OD pairs	
	at distance cutoff					excl. imm.
	6 km	12 km	22 km	32 km	all	OD pairs
e^{RD}	-0.371	-0.448	-0.434	-0.353	-0.387	-0.360
	(0.049)	(0.053)	(0.059)	(0.098)	(0.013)	(0.020)
sample window (km)	[-3,3]	[-3,3]	[-5,5]	[-5,5]	-	-
dist. poly.	1	1	1	1	5	5
N	9354	13393	18307	11560	67571	53175

Note: Estimations from Equation 1 separately at each distance cutoff for Columns 1 through 4, and jointly of all distance cutoffs for Columns 5 and 6. Robust standard errors are in parentheses. The dependent variable is the log change in ridership in the same OD pair between September 2014 and April 2015. All regressions control for the origin and destination fixed effects and are weighted by the ridership in September 2014.

3.1.2 Accounting for Strategic Bunching in Trip Distance

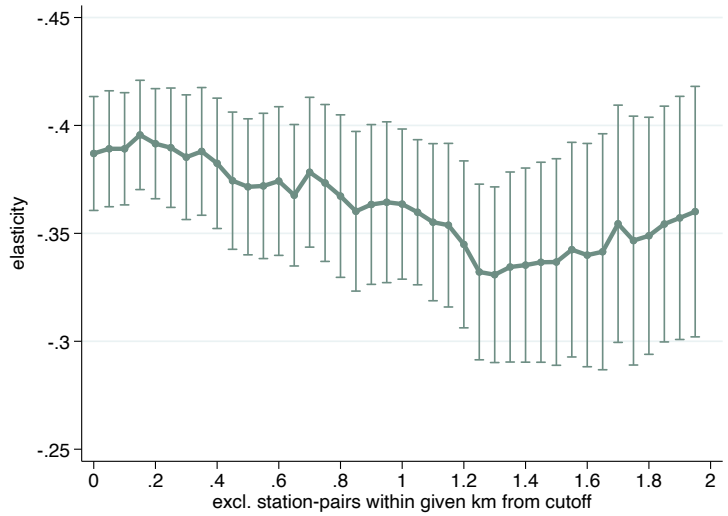
One concern with applying the RD design here is that passengers may strategically bunch below the distance cutoffs. Consider a passenger who would otherwise take a subway trip that has a distance right above the cutoff. She is incentivized to take a slightly shorter ride by starting the ride one station after or ending it one station before. Such bunching behavior represents little change in the actual use of the subway in terms of passenger-miles but would contribute to a

large difference in the number of trips on different sides of the cutoff, as is captured by the RD estimation.

We are interested in a demand elasticity not contaminated by strategic bunching behavior. Our baseline approach is a "donut-hole" RD design (Cattaneo et al., 2019), where OD pairs immediately around the cutoffs are excluded from the regression.[9] Intuitively, although bunching incentives may exist, shortening a subway trip incurs time and pecuniary costs — the passenger may need to walk or bike for a longer distance, take a detour, or ride a bus trip to access the subway or arrive at the final destination. The subway has a clear advantage in speed over those alternative transportation modes, so substantially cutting short the subway ride to save one yuan is unlikely to be worthwhile.

Table 2 Column 6 reports the results from the donut-hole RD estimation. The demand elasticity is about -0.36, which is our preferred estimate of the bunching-free demand elasticity. Consistent with strategic bunching, it is smaller in magnitude than the elasticity using the whole sample (-0.39, Column 5), but only slightly.

Figure 2: Demand Elasticity by Excluding OD Pairs around Cutoffs



Note: The graph shows the point estimator associated with $\Delta \ln P_{od}$ in Equation 1 (which is the demand elasticity) and the associated 95% confidence intervals. The horizontal axis indicates the OD pairs that are excluded from each regression that has distances to the nearest fare threshold within the indicated number of kilometers. Regressions are run for each 50-meter increment in the radius of the hollow-out region.

The donut-hole RD estimation assumes that users do not bunch beyond immediate OD pairs

[9]To implement this approach, we start with all pairs of neighboring stations on the same subway line. Say one such pair of stations is denoted as $(A, A+1)$. We then calculate the distances between this pair and any other station S in the subway system. If the distance between A and S and that between $A+1$ and S lies on either side of a cutoff, OD pairs of (A, S) and $(A+1, S)$, as well as OD pairs of (S, A) and $(S, A+1)$, are excluded from the sample. About 20% of the OD pairs are dropped as a result of this restriction.

around the distance cutoff. To test this assumption, we re-estimate Equation 1 by varying the size of the donut hole. As the distance window gradually widens, it becomes increasingly costly for passengers to bunch. We expect the magnitude of the estimated demand elasticity to decrease with the size of the hole but then stabilize as no one would bunch beyond a certain distance.

Figure 2 summarizes those estimates. The estimated elasticity gradually decreases in magnitude as the donut hole widens, reaches a minimum magnitude of around -0.34 when the donut hole is about 1.2 km wide in radius (2.4 km in diameter), and largely stabilizes around -0.35 afterward. The median distance between adjacent stations is about one km. A 2.4-km window would exclude more than 90% of the station pairs immediately around distance cutoffs and many farther apart. Overall, changes in the estimated demand elasticity are small as the window widens. For estimates with a radius larger than 1 km, we cannot reject that the estimates are statistically the same as that in the baseline (Table 2 Column 6). While the estimation becomes slightly less precise as the donut hole widens and more OD pairs are dropped from the estimation, throughout, the 95% confidence intervals are tightly bounded between -0.4 and -0.3.

We devise alternative approaches to account for strategic bunching in trip distance. Appendix B.2 introduces an approach that directly estimates the "bunching elasticity," which describes the degree to which users are willing to shorten their trips and bunch below the distance cutoff. We estimate a small bunching elasticity of -0.02. Using that to adjust the demand elasticity, we get a demand elasticity that is very close to the baseline. Appendix B.3 introduces a method that exploits the fact that we can link users across different sample periods. The demand elasticity estimated this way is around -0.27. We show that strategic bunching does not affect user-level estimates but are likely a lower bound.

3.2 User Types and Heterogeneity

3.2.1 Classifying User Types

The smartcard data has rich information about trips. Our empirical strategy is also flexible enough to estimate heterogeneous responses by trip types by time, location, and distance. One important limitation of the data is its lack of *user*-level information. The card is anonymous, and no demographic information is associated with it. While aggregate and trip-level analyses can inform how the fare adjustment affects the overall ridership and its composition, they do not speak to distributional impacts. Distributional impacts are important because we ultimately care about *who* are affected and in what ways.

In lieu of demographic information, we classify user (i.e., card) types according to their travel patterns. That task is achieved by adopting a simple machine learning algorithm called the $K-$means clustering. With a predetermined number of clusters (the K), the algorithm groups

users into clusters to minimize within-cluster distances in terms of chosen predictors. The value of K can be cross-validated by inspecting the patterns of clusters. We stop increasing the number of clusters when allowing for a larger K does not lead to a new cluster with a unique pattern.

We apply this algorithm to the 12 million users in the April 2015 data.[10] We first identify *infrequent users*, defined as those who had less than four trips during the month. These users have too few trips to describe their travel patterns. This group, with 5.4 million users, accounts for 40% of all users but only 8% of all trips. On average, an infrequent user took 1.84 subway trips in the month. Three-quarters of their trips occured on weekends or during weekdays' off-peak hours.

Three sets of predictors are used to classify the remaining users. The first set includes three variables that capture the *intensity* of subway use: the total number of trips, the number of weekdays traveled, and the average trip distance. The second set of three variables captures the *timing* of the trips: the share of trips during the weekday morning peak hours, the share during the weekday afternoon peak hours, and the share on weekends. The final set of two variables captures the *geographic* patterns of the trips: the Herfindahl–Hirschman index (HHI) in terms of origin-destination location bins and a measure of the OD location bin concentration rate, which is the total number of trips the user take during the month divided by the number of unique location bin pairs in those trips. Appendix C provides details of the clustering algorithm.

The clustering algorithm classifies four user groups. According to the most salient travel patterns, we name them as follows: weekenders (1.98 million), weekday non-rushers (2.74), rush-hour commuters (1.3 million), and other frequent but less regular commuters (0.96 million). Table 3 summarizes each group's travel characteristics. On average, weekenders take eight subway rides in the month, over 60% of which are on weekends. Weekday non-rushers have an average of nine subway trips, and over half of those are during off-peak hours on weekdays.

Both rush-hour and less-regular commuters are frequent subway riders. The average monthly number of trips for the former group is 25 for the former group and 42 for the latter. The main difference between the two groups is that regular commuters use the subway predominantly for daily commutes. Their trips are concentrated in weekday peak hours (accounting for 80% of all their trips) and have a regular geographic pattern (the location bin HHI is 0.72). Less-regular commuters also use subway for commuting—peak-hour travel accounts for half of their subway rides. But they also use the subway during other times and likely for other purposes. As a result, their trips are less geographically concentrated. All five groups have a similar average trip length of around 15 km.

Appendix Figure C.1 shows the distribution of trips by user type and time. The average rid-

[10] April 2015 is the only month we have full-month ridership data, so we cannot describe users' travel patterns over a longer period. We only have one week of data before the fare adjustment.

Table 3: Card Classification and Characteristics

	infreq. users	weekenders	weekday non-rushers	rush-hour commuters	less-reg. commuters
	(1)	(5)	(4)	(2)	(3)
# of users (mil.)	5.40	1.98	2.74	1.30	0.96
# of rides (monthly)	1.84	8.20	9.16	24.68	42.46
	(0.75)	(4.67)	(5.81)	(13.28)	(13.18)
total distance (km)	29.74	131.89	141.50	379.20	665.99
	(22.13)	(99.19)	(115.60)	(311.69)	(351.88)
share of rides during					
weekday AM rush	0.13	0.07	0.16	0.45	0.28
	(0.28)	(0.11)	(0.16)	(0.19)	(0.14)
weekday PM rush	0.14	0.10	0.17	0.35	0.21
	(0.29)	(0.13)	(0.16)	(0.19)	(0.12)
weekday non-rush	0.39	0.20	0.54	0.14	0.29
	(0.42)	(0.16)	(0.20)	(0.14)	(0.18)
weekend	0.34	0.63	0.13	0.06	0.22
	(0.44)	(0.21)	(0.13)	(0.08)	(0.10)
# of weekdays traveled	0.93	2.31	4.68	13.41	16.98
	(0.72)	(2.00)	(2.81)	(6.11)	(3.64)
location bin HHI	0.82	0.37	0.38	0.72	0.48
	(0.25)	(0.22)	(0.22)	(0.22)	(0.24)
OD location bin concen. rate	1.36	2.14	2.26	9.24	7.47
	(0.52)	(1.36)	(1.43)	(7.81)	(7.06)

Note: Cards are classified into five groups based on their travel patterns in the month of April 2015 using a $K-$means clustering algorithm. The table reports the summary statistics of travel patterns for each card category. See Appendix C for details of the clustering algorithm. Also see Appendix Figure C.1 for the distribution of trips by card type and time.

ership on a weekday is 50% higher than on a weekend. Rush-hour and less-regular commuters account for 65% of the weekday ridership and over 80% of the peak-hour system load.

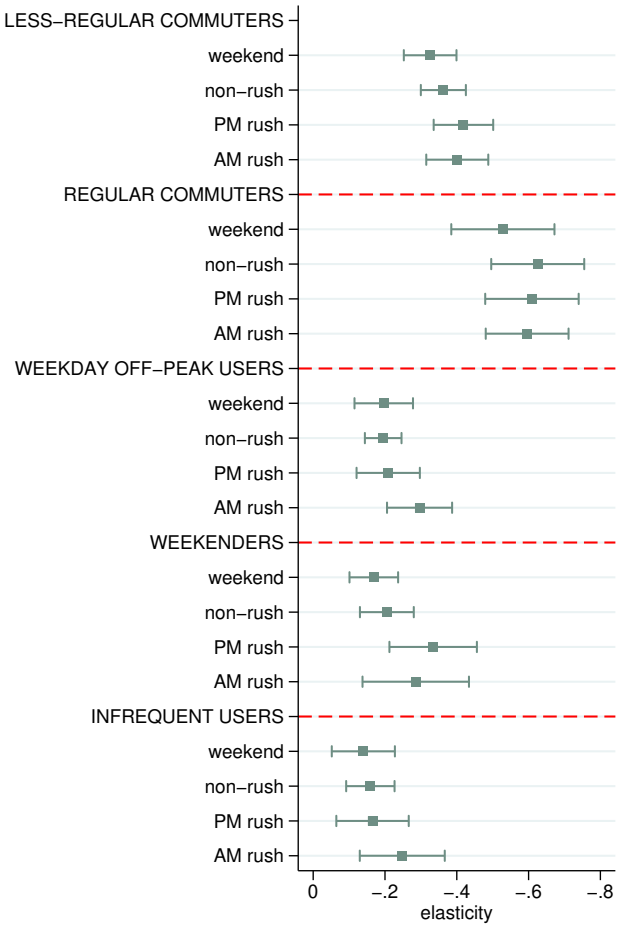
3.2.2 Heterogeneity in Demand Elasticity

We estimate the heterogeneous demand elasticity by user type and time of travel. The combination of user type and travel time is suggestive of the *purpose* of travel. For example, the trip is most likely a commute if a regular commuter takes it during peak hours in her usual OD location bin pairs.

We estimate Equation 1 where we replace the dependent variable with $\Delta \ln(N_{od}^{kh})$, which is the log difference between the ridership in the OD pair by user-type k during time h in April 2015

and the ridership in the same OD pair during the same hour h by *all* users in September 2014.[11]

Figure 3: Heterogeneous Effect by Card Type by Time



Note: The graph reports the demand elasticity and the associated 95% confidence intervals by card type and by time. The log change in the ridership is taken between the ridership in the specific card type and time in April 2015 and the ridership in the corresponding time in September 2014. Each coefficient is estimated using Equation 1 and a specification as in Table 2 Column 6.

Figure 3 summarizes the heterogeneity in demand elasticity. All coefficients are precisely estimated and statistically different from zero at the 95% confidence level. One salient pattern is that the demand elasticity of regular commuters is about -0.55, which is substantially larger than that of other groups. This is a somewhat surprising finding because one may expect regular commuting trips to be difficult to substitute. On the other hand, this group has a strong incentive to replace subway rides with other modes of transportation because they can save large amounts

[11] We do not restrict the comparison to be made within the same user type because the same user's travel patterns may have changed due to the fare adjustment. Nevertheless, classifying users in September 2014 using the same set of predictors yields strikingly similar patterns and distributions of user types. Appendix C provides more details. Defining $\Delta \ln(N_{od}^{kh})$ based on the same type of users results in quantitatively similar results.

of expenses by doing so. They may also face a lower switching cost as their travel patterns are highly predictable. Consistent with this hypothesis, we find that less-regular commuters, who also spend a lot on the subway, have a demand elasticity of around -0.4, which is larger (in magnitude) than that of less frequent users but notably smaller than that of regular commuters.

The larger demand elasticity among commuters has implications for the distributional impacts of the fare adjustment and externality on road congestion. Changes of work or home locations due to the subway fare rise are likely of second order, at least in the short run, and workers need to fulfill their commutes using other modes of transportation, which likely generate a larger congestion externality than subway trips.

Another pattern from Figure 3 is that the demand elasticity differs by user type, but within the user type, it does not substantially vary by the type of trip. Regular commuters have a higher demand elasticity, and that elasticity carries over not only to commuting trips during peak hours, but also to trips during non-peak hours and on weekends that are probably not work-related. We cannot reject that demand elasticities within the same user type are statistically identical. This also suggests that users do not cross-substitute between trips of different types and for different purposes. For example, it is possible that commuters would disproportionately cut non-commuting trips and preserve commuting trips in response to the fare rise. If that is the case, we would expect a larger demand elasticity for the former and a smaller elasticity for the latter. But this is not what we find here. Therefore, in counterfactual exercises, we assign each type of user a single elasticity.

3.3 Behavioral Responses to Cumulative Quantity Discounts

3.3.1 The Rational, the Myopic, and the Oblivious

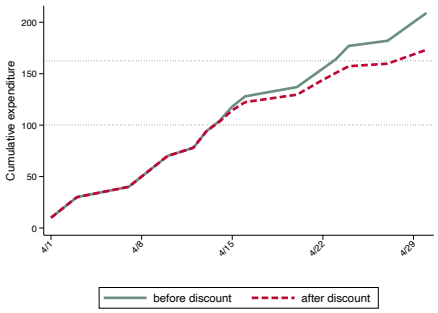
The analyses so far have exploited discontinuities in the *listing* price created by the new fare structure. However, the progressive discount schedule creates kinks in the monthly budget and potentially makes the *out-of-pocket* price different from the listing price. Understanding how consumers respond to the fare adjustment depends crucially on what price they respond to.

Figure 4 illustrates the issue using one user's travel records in April 2015. Panel A shows the user's cumulative expenditure by the date of the month. The solid green line shows the expenditure according to the listing price, while the red dashed line shows the out-of-pocket expenditure after applying discounts. The user pays the listing price until the month-to-date cumulative expenditure surpasses the first threshold at 100 yuan. After that, she pays 80% of the listing price until her out-of-pocket expenditure reaches 150 yuan, after which she pays only 50% of the listing price. In that month, the user eventually had trips that were worth 209 yuan, for which she paid only 173.25 yuan. By the end of the month, she faced a *marginal* discount rate of

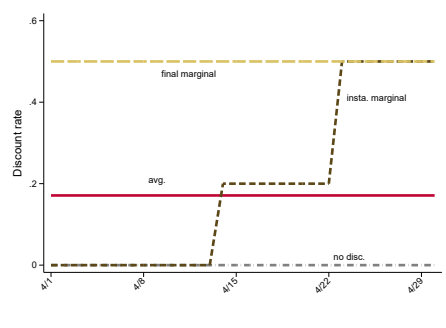
50%. Overall, she enjoyed an *average* discount rate of 17% on her trips during that month.

Figure 4: Rational, Myopic, and Oblivious: An Example

Panel A: Subway Expenditure by Date



Panel B: Four Perceived Discount Rates



Note: The graphs illustrate month-to-date expenditure and associated prices using one user observed in April 2015. Panel A shows cumulative expenditure by date. The green solid line shows cumulative expenditure before discounts, and the red dashed line shows out-of-pocket expenditure. Panel B illustrates the four different perceptions of discount rates the user may respond to.

Panel B shows the paths of different discount rates to which the user may respond. A fully forward-looking and *rational* consumer responds to the *end-of-month marginal* discount rate, which is 50%. It is important to note that it is the marginal discount rate the consumer faces regardless of the date she plans to take an additional trip. Even if she decides to take another trip at the beginning of the month, although that trip itself may not qualify for any discount, it would "bump" a later trip into the 50%-off region.

Being rational requires a user to fully predict her demand for subway trips for the entire month and correctly react to the marginal price. Instead, the user may respond to the month-*average* discount rate. Responding to the average price instead of the marginal price is a common behavioral bias (Liebman and Zeckhauser, 2004). We call the user an *ironer* as if she "irons" the kinked budget flat and smooth.

Alternatively, the user may base her decision on the *instantaneous marginal* discount rate applied to the current trip. We call such a user *myopic* because she misses the point that the current discount rate results from her past trips, and the trip she takes today has implications for the discount rate she would receive for future trips. As illustrated in Figure 4, the instantaneous marginal discount rate a consumer faces may change over the course of a calendar month.

Finally, we say the user is *oblivious* to discounts if she responds only to the listing price even if she actually qualifies for some discount.

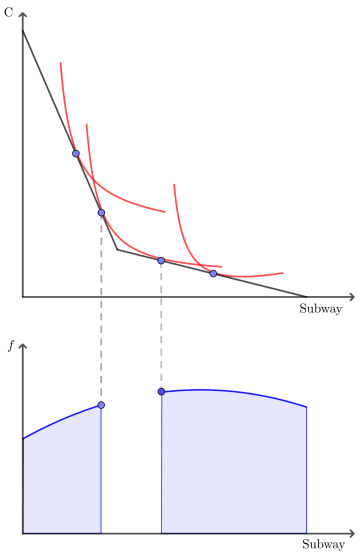
There is a large literature documenting that consumers often fail to respond optimally to non-linear pricing (Liebman, 1998). The sub-optimal choices may be due to consumers' inability to fully grasp the incentives created by the non-linear pricing (Ito, 2014), or their inability to flexibly adjust their choices (Borenstein, 2009; Saez, 2010). The salience of the pricing schedule

has also been shown to be an important factor (Chetty et al., 2009). Imperfect rationality, costs in adjustment, and salience are all likely present in the case under study. This subsection aims to empirically identify the composition of consumer behavioral types that best fit the observed choices.

3.3.2 A Test for Rationality

We start with a test for consumers being rational. The discount schedule creates kinks in the monthly budget constraint, which we use to identify the demand elasticity using the bunching estimator (Saez, 2010; Kleven, 2016).[12]

Figure 5: Kinked Budget Constraint and Demand Elasticity



Note: The graph illustrates the hole in the distribution of consumption of subway trips when there is a non-convex kink in the monthly budget constraint. The x-axis indicates the pre-discount monthly expenditure on the subway with the new fare structure. The y-axis of the graph above, C, indicates the consumption of the numeraire good. The y-axis of the graph below, f indicates the density of users with the corresponding pre-discount monthly expenditure on subway

Figure 5 illustrates the intuition of the test. To start simple, we assume consumers are governed by the same demand elasticity but differ in their idiosyncratic preferences for subway trips. In the graphical illustration, it means that indifference curves from different consumers have the same shape and curvature but differ in their locus. We further assume those idiosyncratic preferences follow a smooth distribution. With a linear budget line, observed consumption of subway trips (using monthly pre-discount expenditure on subway rides as a proxy) will also follow a

[12]The discount schedule creates two non-convex kinks (at 100 yuan and 162.5 yuan before discounts) and one convex kink (at 662.5 yuan before discounts). We focus on the first two non-convex kinks because there are few users around the neighborhoods of the third kink.

smooth distribution. A non-convex kink in the budget, on the other hand, creates a subset of subway consumption values that are strongly dominated. No rational, utility-maximizing users would choose a value inside that region. This creates a hole in the distribution of subway consumption around the kink point, as illustrated by the bottom graph of Figure 5.

The width of the hole is indicative of the demand elasticity. To see that, consider the extreme case where preferences are Leontief. In this case, general consumption and subway trips are perfect complements, and the demand elasticity is zero. The non-convex kink and its surrounding region are not strongly dominated, and we will observe no hole in the distribution. In contrast, if general consumption and subway trips are close substitutes, the indifference curve is close to linear, and the demand elasticity is large. The non-convex kink will lead to a large hole in the distribution.

In reality, we may not see a strict hole in the distribution around a non-convex kink. There are two main reasons for that. First, consumers may be heterogeneous in the *shape* of their indifference curves, corresponding to differences in the demand elasticity. In this case, the hole becomes a "bowl." The kink point remains weakly dominated, so the bowl touches zero as long as the number of consumers with a Leontief preference has measure zero. The width of the bowl at its mouth still allows for the estimation of the *average* demand elasticity. Second, consumers may face friction in optimizing or adjusting their behavior to achieve the optimal point. Larger optimization friction makes the bowl shallower, but the kink point is still likely the bowl's lowest point. With additional assumptions on the distributions of preference and optimization frictions, one can still recover the average elasticity from the width and the depth of the bowl (Kleven, 2016).

Figure 6 shows the empirical distribution of users by their pre-discount monthly expenditure on subway trips. Each graph plots the 30-yuan window around each of the non-convex kink points. The density of users appears to be smooth in the neighborhoods of the kink points. There is no visual evidence of a hole, a bowl, or even a dent in the distribution. The locally weighted scatter-plot smoothing (LOWESS) lines (blue solid lines) also show no sign of a hollowing region around the kink points.

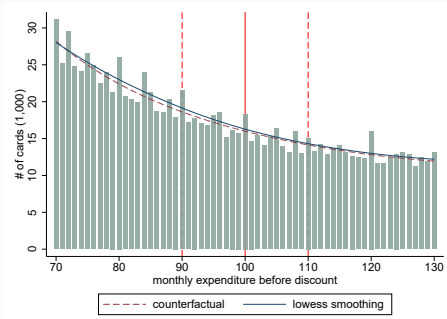
To formally test whether there is any non-smoothness in the distribution, we estimate the following equation:

$$N_j = \sum_{p=0}^{P} \beta_p \cdot Q_j^p + \sum_{i=Q_1^{sub}}^{Q_2^{sub}} \gamma_i \cdot 1(Q_j = i) + v_j. \tag{2}$$

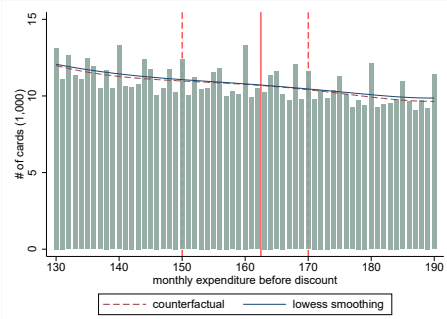
N_j is the number of users who spent Q_j yuan before discounts (in integers) during the month. $\sum_{p=0}^{P} \beta_p \cdot Q_j^p$ is a polynomial of Q_j, which fits a smooth function of the distribution. We then

Figure 6: Density Distribution around Non-convex Kinks

Panel A: First Kink at 100 *yuan*



Panel B1: Second Kink at 162.5 *yuan*



Note: Data are from the full-month ridership records in April 2015. The discount schedule creates three kinks in the budget line, here we show the distribution of pre-discount expenditure at the first two kinks at 100 and 162.5 yuan, respectively. In both graphs, the solid vertical line indicates the kink point, the two dashed vertical lines indicate the neighborhood that is excluded when we impute the counterfactual density. The dashed red line depicts the counterfactual distribution fitted by a polynomial excluding the neighborhood around the kink point. The blue line depicts the smooth fitted line with the neighborhood included. Fitted density and actual density is imputed from estimating Equation 2. The non-convex kinked budget would imply the actual distribution (green bars) to be below the counterfactual distribution in the narrow neighborhood around the kink point ($\hat{\gamma}_j < 0$). In Panel A, the joint test of $\hat{\gamma}_j$ for j between the neighborhood has a $p-$value of 0.98, the summation of those $\hat{\gamma}_j$'s is 0.168, while the average N_j within this range is 16.19. Note that the ratio, 0.168/20/16.19, is a very small number. In Panel B, the joint test of $\hat{\gamma}_j$ for j between the neighborhood has a $p-$value of 0.5, the summation of those $\hat{\gamma}_j$'s is 3.24, while the average N_j within this range is 10.88.

allow the density in the close neighborhood around the kink point to deviate flexibly from the smoothed line. This is captured by $\sum_{i=Q_1^{sub}}^{Q_2^{sub}} \gamma_i \cdot 1(Q_j = i)$, where γ_i captures the magnitude of deviation at $i = j$. For the baseline, we pick $[Q_1^{sub}, Q_2^{sub}]$ to be a 20-yuan window around the kink point (indicated by the two red vertical dashed lines in Figure 6).

The smoothed counterfactual density can be recovered by $\hat{N}_j = \sum_{p=0}^{P} \hat{\beta}_p \cdot Q_j^p$, which is plotted in red dashed lines in the graphs. Notice that they fit closely with the density bars and the LOWESS lines. If there is a hollowing-out region around the kink, we expect $\hat{\gamma}_j < 0$. However, in the neighborhoods of both kink points, $\hat{\gamma}_j$'s are all individually and jointly small and statistically insignificant. Around the first kink, the joint test of $\hat{\gamma}_j$ being statistically different from zero has a $p-$value of 0.98, the summation of $\hat{\gamma}_j$ is 0.168, while the average N_j within this range is 16.19. The summation not only has the "wrong" sign, but it only accounts for 0.05% of the average density in the region (0.168/20/16.19). $\hat{\gamma}_j$'s are also small and statistically insignificant around the second kink.

The lack of a hollowing-out region is not because consumers all have a Leontiff preference. In fact, we show in Section 3.1 clear evidence that users respond to the *listing* price. Therefore, we can rule out that users are strongly or approximately rational.[13] The sharp responses to differ-

[13]Notice that one needs to be a reasonably frequent user to be in the neighborhood of the first kink. One may

ences in the listing prices also suggest that it is unlikely users face exorbitantly high adjustment costs.

3.3.3 A Simple Test for Myopia

If users do not respond to the month-end marginal discount, do they respond to the *instantaneous* marginal discount? We devise a simple test for myopia. The test is embedded in the baseline OD pair RD design and builds on the intuition that users who consistently take trips in OD pairs that are just above a distance cutoff are more likely to qualify for larger discounts than those who travel in OD pairs that are just below the cutoff. Therefore, on a given day, trips in OD pairs right above the cutoff receive a larger marginal discount rate *on average*. Furthermore, given the progressive design of the discount schedule, trips in above-the-cutoff pairs likely receive *increasingly* larger discounts compared with those in below-the-cutoff pairs. There will also be discontinuities in prices around the cutoff if users are rational, ironing, or oblivious, but those discontinuities do not change during the course of the month. The different temporal patterns in price discontinuity thus allow us to test whether consumers are myopic. Intuitively, if we estimate the demand elasticity using the listing price while users are in fact myopic, we would see the estimated demand elasticity change over the course of the month as the listing price increasingly misrepresents the actual price users perceive and respond to.

We first demonstrate that discontinuities in different prices do evolve over time, as expected. We run the following regression, which is modified from Equation 1:

$$\ln(p_{od,t}^{\text{price type}}) = \rho_t^{\text{price type}} \cdot \ln p_{od}^{\text{listing}} + g(dist_{od}) + \Phi_{o,d,t} + \varepsilon_{odt}. \tag{3}$$

Each observation is an OD-pair-by-date in April 2015. Price type is one of the following: listing price, price after applying the instantaneous marginal discount (inst mar), price after applying the end-of-month marginal discount (mar), and price after applying the average monthly discount (avg). $p_{od,t}^{\text{price type}}$ is the average price of trips in OD pair od in day t according to the specific behavioral response.

Equation 3 is separately run for each Monday through Thursday in April 2015.[14] $\rho_t^{\text{price type}}$ captures the discontinuity in the specific price around distance cutoffs. ρ_t^{listing} is equal to one by

suspect that frequent users with a regular travel pattern are more likely to predict their monthly demand and rationally respond to it. Appendix Figure B.5 shows that there is no visual or statistical evidence of hollows or dents around the first two nonconvex kinks in the monthly budget among users we classify as regular commuters.

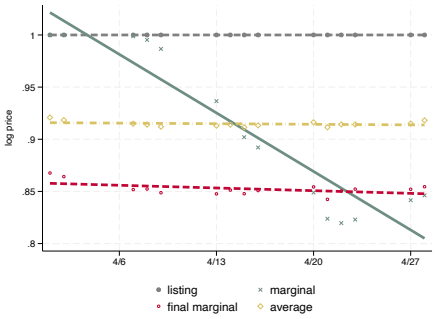
[14]The average final marginal discount and the average discount of trips in an OD pair are similar over different days only when the composition of users who take these trips remains similar throughout the month. The compositions of users in a given OD pair are substantially different between weekdays and weekends. For the same reason, we also drop one national holiday that landed in the month, as well as Fridays, which apparently have substantive leisure trips and have a user composition that is different from those between Monday and Thursday.

definition. We expect $0 < \rho_t < 1$ under other price types, while ρ_t^{mar} and ρ_t^{avg} are expected to be constant over time, $\rho_t^{\text{inst mar}}$ is expected to decrease over time.

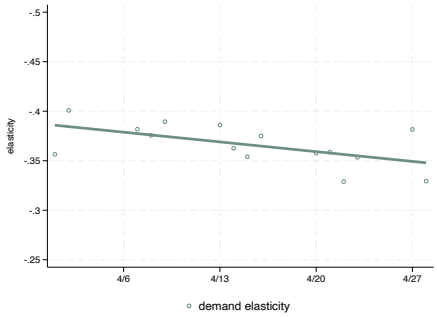
Panel A of Figure 7 plots $\rho_t^{\text{price type}}$. Except for ρ_t^{listing}, they are all below one. $\rho_t^{\text{inst mar}}$ and ρ_t^{avg} are largely constant over t. Mild fluctuations around the flat fitted lines reflect small day-to-day differences in the composition of users. In contrast, $\rho_t^{\text{inst mar}}$ declines over time. In fact, the path of $\rho_t^{\text{inst mar}}$ follows a flipped $S-$shape. It is close to one at the beginning of the month when no one is yet qualified for any discount. Then those who take slightly longer but discontinuously more expensive trips start to qualify for some discounts, leading to a declining $\rho_t^{\text{inst mar}}$. Towards the end of the month, those who take slightly longer trips hit a plateau in the instantaneous marginal discount rate, while those who take slightly shorter trips gradually catch up. $\rho_t^{\text{inst mar}}$ slightly increases and then flattens out.

Figure 7: A Test for Myopia

Panel A: RD in Prices



Panel B: $\hat{e}_t^{\text{listing}}$



Note: Panel A plots regression discontinuity estimates of perceived prices under various behavioral assumptions by each day between Monday and Thursday in the month of April 2015. The regression equation is described in Equation 3. The sample includes all trips. Coefficients associated with each estimation are plotted, and lines in the corresponding color are linear fits of those coefficients. The linear fitted line for $\hat{\rho}_t^{\text{myopic}}$ has a slope of -0.008 and a robust standard error of 0.0009. Panel B plots regression discontinuity estimates of demand elasticity by date in the month of April 2015, assuming consumers respond only to the listing price. The regression equation is described in Equation 4. The sample includes all trips. The linear fitted line has a coefficient of 0.0014 and a robust standard error of 0.0007. Confidence intervals are suppressed in both panels for clean illustration.

Now consider estimating the following equation separately for each day t:

$$\Delta \ln(N_{od,t}) = e_t^{\text{listing}} \cdot \ln p_{od}^{\text{listing}} + g(dist_{od}) + \Phi_{o,d,t} + \Delta\varepsilon_{odt}. \tag{4}$$

$\Delta \ln(N_{odt})$ is the log change in ridership in an OD pair between day t in April 2015 and the corresponding day of the week in September 2014. p_{od}^{listing} is the listing price. e^{listing} is the demand elasticity associated with the listing price as Equation 4 is estimated separately for each date t.

We assume the *true* demand elasticity to be a constant, which means that users always respond to whatever price they perceive in the same way, regardless of the day of the month. If consumers

are oblivious, rational, or ironing, $\hat{e}_t^{\text{listing}}$ will be constant over t. In contrast, if consumers are myopic, $\hat{e}_t^{\text{listing}}$ will increasingly *under*-estimate the true elasticity because the discontinuity in the listing price increasingly *over*states the discontinuity in the actual perceived price.

$\hat{e}_t^{\text{listing}}$ are plotted in Figure 7 Panel B. The fitted linear line is slightly downward trended, and we can reject that the slope is zero at 10% statistical level (the slope is 0.0014, with a robust standard error of 0.0007; note that the y-axis is reversely labeled). This is consistent with myopia. However, the magnitude of the downward trend is small. Panel A shows that the discontinuity in price declines by 24% over the course of the month. If users are fully myopic, this would imply $\hat{e}_t^{\text{listing}}$ by the end of the month to be 24% smaller than that at the start of the month. Panel B shows that the actual decline is about 11%.[15]

3.3.4 Estimating the Composition of Behavioral Types

The previous two subsections show that users are definitely not rational, nor are they overwhelmingly myopic. Realistically, users may differ in sophistication and in prices they respond to. This subsection aims at estimating the composition of behavioral types in the user population. We run a mixture model where all four prices to which different behavioral types respond are included in the OD-pair RD framework. The model determines shares of different behavioral types that best fit the data. Consider the following equation:

$$\begin{aligned}\Delta\ln(N_{od,t}) = {} & \beta^{\text{listing}}\ln(p_{od}^{\text{listing}}) + \beta^{\text{inst mar}}\ln(p_{od,t}^{\text{inst mar}}) + \beta^{\text{avg}}\ln(p_{od,t}^{\text{avg}}) + \beta^{\text{mar}}\ln(p_{od,t}^{\text{mar}}) \\ & + f(X_{od}) + \Phi_{o,d} + \lambda_t + \varepsilon_{od}. \end{aligned} \tag{5}$$

Each observation is an OD-pair-by-date. Here, we include all dates in April 2015 and estimate the model in one regression (pooling all dates). $\Delta\ln(N_{od,t})$ is the log change in the number of trips in the OD pair between date t of April 2015 and the corresponding day of the week in September 2014. The four prices that users may potentially respond to are in log forms and are averaged across trips in OD-pair od in day t. $f(X_{od})$ is a flexible function of the OD-pair distance, for which we use a flexible polynomial up to the 5^{th} order. $\Phi_{o,d}$ is the origin and destination station fixed effects. λ_t is the date fixed effect. The remainder of the specification is the same as in Column 6 of Table 2. Standard errors are clustered at the OD-pair level.

The variation used to separately identify each price comes from two sources. First, given each date, OD pairs on different sides of the cutoff have different price discontinuities that vary by the price type. As illustrated by Figure 7 Panel A, the discontinuity is largest in the listing price and is typically the smallest in the month-end marginal price. Second, for all price types except for

[15] Appendix Section B.5 presents additional evidence using only frequent users and shows that myopic behavior is limited.

the listing price, the average price in a given OD pair changes over time, either because the same user faces different instantaneous marginal prices over the course of the month or because the user composition in the same OD pair varies by date.[16]

β's contain two pieces of information. First, they indicate how users respond to price, captured by the demand elasticity. Second, they show what fraction of users respond to *each* price, thus indicating the mixture of behavioral types. To see that, note that Equation 5 can be rewritten in the following form:

$$\begin{aligned}\Delta\ln(N_{od,t}) = {} & e \cdot \gamma^{\text{listing}}\ln(p_{od}^{\text{listing}}) + e \cdot \gamma^{\text{inst mar}}\ln(p_{od,t}^{\text{inst mar}}) + e \cdot \gamma^{\text{avg}}\ln(p_{od,t}^{\text{avg}}) \\ & + e \cdot (1 - \gamma^{\text{listing}} - \gamma^{\text{inst mar}} - \gamma^{\text{avg}})\ln(p_{od,t}^{\text{mar}}) \\ & + f(X_{od}) + \Phi_{o,d} + \lambda_t + \varepsilon_{od}. \end{aligned} \tag{6}$$

e is the demand elasticity. γ's are the composition of users who respond to various prices. They add up to one across price types. From reduced form estimates of β's, we can jointly recover structural parameters e and γ's.

An implicit assumption imposed here is that the demand elasticity is a constant. The model is unidentified if e is heterogeneous, and the heterogeneity is correlated with user behavior types — for example, if regular commuters, who have a higher elasticity, are also more likely to be rational. To alleviate this concern, we also estimate the model using more homogeneous subgroups. We consider two such sub-samples: (1) a group of frequent users with a pre-discount monthly expenditure greater than 70 yuan and (2) a further refined group of frequent users who have regular travel patterns (regular commuters). Those are also users who are likely to qualify for discounts, such that distinguishing the four different prices is relevant.[17]

Users may face prediction and optimization frictions even if they intend to respond to the non-linear budget. We introduce a flexible error term for the end-of-month marginal price and the monthly average price to capture imperfect optimization. Equation 6 can be further extended as:

$$\Delta\ln(N_{od,t}) = e \cdot \gamma^{\text{listing}}\ln(p_{od}^{\text{listing}}) + e \cdot \gamma^{\text{inst mar}}\ln(p_{od,t}^{\text{inst mar}})$$

[16]Note that we include all days in April 2015 in the regression. The user compositions in a given OD pair are substantially different between weekdays and weekends, which lends day-to-day variation in the month-end marginal and average prices. Consequently, except for the listing price, all other prices contain a t subscript in Equation 5. Nevertheless, it remains true that the four prices are highly correlated, and in some cases, there may not be sufficient statistical power to identify them separately. We thus also estimate the model using three-price or two-price mixtures.

[17]The choice of 70 yuan as the cutoff is arbitrary. We want to include users who would expect to qualify for some discount and potentially respond to it. This includes users who actually qualified for discounts but also those who expected to but eventually did not. The exact choice of the cutoff value is inconsequential. The results are quantitatively similar to a wide range of choices of the cutoff value.

$$
\begin{aligned}
&+ e \cdot \gamma^{\text{avg}}[\ln(p^{\text{avg}}_{od,t}) + g_1(\ln(p^{\text{avg}}_{od,t}))] \\
&+ e \cdot (1 - \gamma^{\text{listing}} - \gamma^{\text{inst mar}} - \gamma^{\text{avg}})[\ln(p^{\text{mar}}_{od,t}) + g_2(\ln(p^{\text{mar}}_{od,t}))] \\
&+ f(X_{od}) + \Phi_{o,d} + \lambda_t + \varepsilon_{od}. \qquad (7)
\end{aligned}
$$

$g_1(.)$ and $g_2(.)$ are flexible functions of the corresponding log prices. We proxy them with flexible polynomials up to the 5^{th} order. Equation 7 is under-identified without further restrictions. We allow for only one term with optimization errors at each time and impose e to be the estimated demand elasticity from the corresponding model where prediction errors are not included.

Finally, all discounted prices are functions of ridership. They are thus mechanically correlated with idiosyncratic daily shocks to demand. To break the simultaneity, we construct predicted discount rates by replacing each user's ridership of the day with the average ridership on other days on the same day of the week during the month.[18]

Table 4 Column 1 estimates Equation 5 using trips from all users and includes all four price types. Summing over the β's, the results suggest a demand elasticity of -0.351, which is close to the baseline estimate. According to this estimation, oblivious users account for 93% of the trips.[19] However, there is some evidence that the model does not have sufficient variation to identify all four prices separately. The coefficients have large standard errors and are not statistically significant.

Column 2 reports the same regression using trips from frequent users, for whom discounts are more likely to be relevant. The implied demand elasticity is -0.551. The regression shows that frequent users are also predominantly oblivious to the discounts. In addition, the coefficient associated with the log average price has the "wrong" sign, which may indicate it being a poor behavioral assumption.

Columns 3 and 4 report estimation results from three-price mixture models in which the end-of-month marginal and monthly average prices are included separately. With three-price mixtures, the models are estimated more precisely. The implied demand elasticity in the two models, -0.569 and -0.560, remain similar to that in Column 2. In both columns, the entirety of the elasticity is loaded on the listing price. Coefficients associated with other prices are small in magnitude and not statistically significant.

Column 5 reports the results from a re-estimation of the model in Column 3 allowing for prediction errors in the end-of-month marginal price. γ^{marginal} is not identified, while γ^{listing} and $\gamma^{\text{inst marginal}}$ are identified by restricting the demand elasticity to that estimated from Column 3. Column 6 re-estimates the model in Column 4 while allowing for prediction errors in the

[18] For example, there were five Wednesdays in April 2015. If t is on the first Wednesday of the month, we replace a user's ridership as the average expenditure on the other four Wednesdays of the month.

[19] $\gamma^{\text{listing}} = -0.325/-0.351$. It represents the share of trips because each observation is weighted by the number of trips in the OD pair on the corresponding day in September 2014, not by the number of users.

Table 4: Mixture Model and the Composition of Behavioral Types

	dep var: $\Delta\ln(N_{od,t})$							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Log listing p	-0.325	-0.914	-0.582	-0.635	-0.523	-0.747	-0.790	-0.741
	(0.192)	(0.198)	(0.081)	(0.182)	(0.007)	(0.011)	(0.008)	(0.014)
Log instan. marginal p	0.055	0.003	0.030	0.000	-0.114	-0.140	0.089	0.048
	(0.042)	(0.029)	(0.033)	(0.029)	(0.004)	(0.004)	(0.006)	(0.006)
Log final marginal p	-0.088	-0.196	-0.017		-		-	
	(0.126)	(0.072)	(0.063)		-		-	
Log avg. p	0.007	0.556		0.075		-		-
	(0.284)	(0.239)		(0.184)		-		-
Polynomials of								
log final marginal p					X		X	
log avg. p						X		X
e constrained at					-0.569	-0.560	-0.534	-0.527
Sample	all	frequent					freq. and regular	
# of obs. (mil.)	1.47	1.39					1.20	

Note: The table reports results from estimating various versions of the mixture model. Each observation is an OD pair by date. The dependent variable the log difference between the ridership from the specific sample of users in that OD pair on a day of April 2015 and the ridership from the corresponding user group and on the same day of the week in September 2014. For Column 1, the sample includes ridership from all users. In Columns 2 through 6, the sample includes trips from frequent users. In Columns 7 and 8, the sample includes trips from frequent users with regular commuting patterns. Columns 5 through 8 account for optimization errors by including either a 5^{th}-order polynomial of log final marginal price (Columns 5 and 7) or that of log average price (Columns 6 and 8). In those regressions, the overall elasticity is constrained to be those estimated in the corresponding specifications in which optimization errors are not accounted for. All regressions include a 5^{th}-order polynomial of the OD-pair distance. Log instantaneous marginal price, log final marginal price, and log average price are instrumented using counterparts that replace the same-day actual ridership with the predicted ridership. Standard errors are two-way clustered at the origin and destination stations.

monthly average price. Both estimations still point to the conclusion that users are predominantly oblivious. The same conclusion holds in Columns 7 and 8, where the model is estimated using trips from regular commuters.

Appendix B.6 presents results from additional checks. All evidence suggests that users are predominantly oblivious to cumulative quantity discounts.

3.3.5 Summary of Behavioral Responses to Quantity Discounts

The analyses in this subsection show that users of the Beijing subway do not respond to quantity discounts either rationally or heuristically. This is true even among users who could potentially have benefited greatly from the discounts. While identifying the sources of such suboptimal behavior is beyond the scope of this paper, We discuss some potential explanations here.

The first explanation is that consumers face mental and environmental frictions in calculating and carrying out theoretically optimal actions. Such frictions lead to *approximately* optimal solutions, which we find no evidence for. We also find no evidence that consumers use *heuristic* optimizing rules by responding to the average and instantaneous marginal prices.

The second explanation is the difficulty in predicting demand for the entire month. There are substantial week-to-week fluctuations in subway trips, even among those with relatively regular travel patterns. However, a user who fails to optimize at the monthly level could still respond to the instantaneous marginal price, which, as shown in the next section, still leads to utility gains compared with completely ignoring the discounts. Yet we find no evidence for consumers being myopic either.

The third explanation is that discounts are not salient. During the sample period, there is no easy way to know how much one has spent during the month and what discount she receives in the next few trips. When tapping out of a station, the small screen on the gate shows how much is charged for the trip and the remaining balance on the smartcard. But passengers of Beijing's subway face a hectic environment and are constantly in a rush. It may be difficult to back out the instantaneous discount rate within a split second, and it is even harder to predict the monthly expenditure without diligently keeping a log of subway trips.

Finally, how users actually respond to the non-linear monthly budget has implications for a better design of the fare structure. In Section 4, we show that behavioral responses to the cumulative quantity discounts have implications for both aggregate and distributional welfare. Then we propose alternative fare structures and analyze their welfare implications under various behavioral assumptions.

3.4 Rescheduling Elasticity

Starting in December 2015, 16 stations on two subway lines that were often crowded during morning peak hours adopted an early-bird discount (EBD). Card users entering one of those stations before 7 AM on a weekday receive a 30% discount. The EBD creates a sharp discontinuity in price in the *timing* of travel, and there is salient evidence that users respond to it. Panel A of Figure 8 shows the scatterplot of the total number of users who entered those stations by each minute between 5:30 AM and 9 AM in the week of September 12, 2016 (Monday through Friday). There is an abrupt drop in the number of entries after 7 AM. In contrast, Appendix Figure B.9 Panel A shows that there was no such discontinuity in September 2015, before the EBD was implemented. The discontinuity in ridership allows us to estimate the flexibility of users' travel schedules.

The sharp discontinuity in the number of passengers around the EBD cutoff time comes from

two sources. First, the discount creates a larger demand for trips before the time cutoff. We call trips induced by the lower fare *opt-in* trips, the number of which is governed by the demand elasticity, denoted as e^{d}. Second, the price difference between trips taken before and after the time cutoff creates an incentive to *reschedule* trips planned for after 7 AM to some time before. The number of rescheduled trips is governed by a set of *rescheduling elasticity*, denoted as e_t^{r}, which is defined as the percent of trips originally planned for time t ($t > 7$ AM) that is rescheduled to some time before 7 AM due to a one percentage point increase in the EBD. The rescheduling elasticity is a function of the originally planned time t because, presumably, the cost of rescheduling a trip increases with the difference between the originally planned time and the time qualified for the EBD.

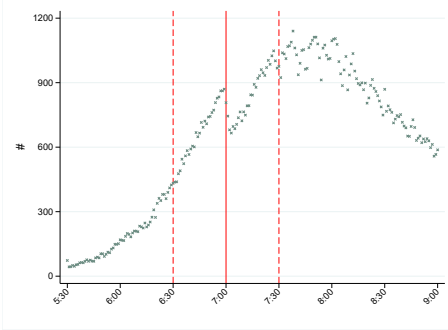
Figure 8 Panel B illustrates the composition of observed entries with the presence of the EBD. The graph shows the actual ridership data and the parameters we estimate below. Assuming without the EBD, the counterfactual number of entries is a smooth function of time. To the left of the cutoff, the observed number of entries is the sum of the counterfactual number of trips (represented by the green bars), opt-in trips (red bars), and trips rescheduled from some time after the cutoff (blue bars and henceforth referred to as the *bunching mass*). To the right of the cutoff, the observed number of entries (green bars) is the difference between the counterfactual trips and those rescheduled to an earlier time (hollow bars and henceforth referred to as the *missing mass*). The missing mass on the right should equal the bunching mass on the left.

The goal here is to estimate the rescheduling elasticity, e_t^{r}. The sample includes the number of entries into the stations with EBD between 5:30 AM (when most subway stations start to operate) and 9 AM on weekdays. In the baseline, we take the demand elasticity from the preferred estimation in the OD-pair RD design (-0.36, from Table 2 Column 6).

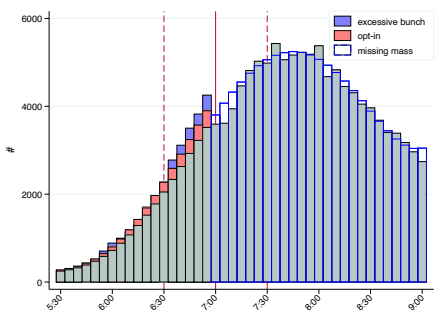
To estimate e_t^r, we need to decide the *rescheduling window* within which changes of travel time take place. We assume only trips originally planned for between 7 and 7:29 AM have the incentive to be rescheduled, and the rescheduled trips will land in the time window between 6:30 and 6:59 AM. The former part of the rescheduling window is referred to as the *missing window* and the latter as the *bunching window*. In theory, users have no incentive to reschedule a trip to any time before 6:59 AM. But in practice, users may be risk-averse or may strategically try to avoid crowding by choosing to arrive a few minutes earlier. A wide bunching window allows for those practical considerations. The widths of the missing and bunching windows can be cross-verified by inspecting whether the missing and bunching masses are bounded within the chosen windows. Ridership in the sample time window but outside the missing and rescheduling windows, adjusted for opt-in trips, are used to fit the curve of counterfactual ridership. Obviously, a wider rescheduling window allows for more flexible rescheduling behavior but cuts a larger hole in the sample and makes the estimation of the counterfactual ridership less precise. We test the

Figure 8: Early Bird Discount and Rescheduling Elasticity

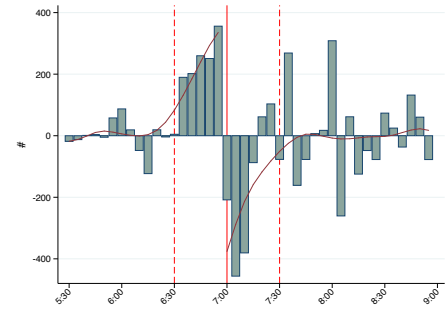
Panel A: Entries by Minute



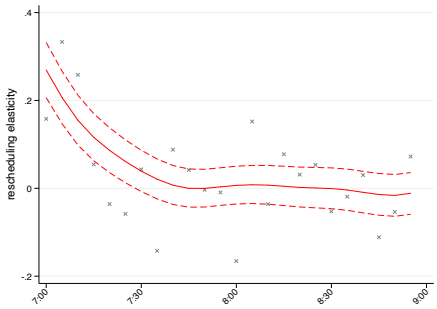
Panel B: Decomposing Entries



Panel C: Missing and Bunching Masses



Panel D: Rescheduling Elasticity



Note: The data include all trips in the week of September 12, 2016. The sample includes trips that start between 5:30 AM and 9 AM and originated from 16 stations that had the EBD implemented in December 2015. Panel A shows the number of trips by the time of entry in those stations. The two verticle red dashed lines indicate the rescheduling window in the baseline. Panel B illustrates the composition of ridership. Before 7 AM, observed trips include trips that would have been taken without the EBD (green bars), opt-in trips due to the EBD (red bars), and trips rescheduled from after 7 AM (blue bars). After 7 AM, observed trips (green bars) are the difference between trips that would have been taken without the EBD and trips rescheduled to some time before 7 AM (hollow bars). Panel C shows the excessive bunching (before 7 AM) and the missing mas (after 7 AM) in five-minute bins. The red curves are non-parametric fit for the size of the missing mass and that of the bunching mass, respectively. Panel D reports the associated rescheduling elasticity by five-minute bins. Smoothed 95% confidence intervals (the smoothed 2.5^{th} percentile and the smoothed 97.5^{th} percentile from 1,000 bootstraps) are in dashed lines.

robustness of the results with alternative rescheduling windows.

The estimation follows several steps.

Step 1. Let N_t be the observed number of entries at time t and N_t^c be the counterfactual ridership. Entries to the left of the rescheduling window (between 5:30 and 6:29 AM) consist of counterfactual trips and opt-in trips. With imposed demand elasticity, counterfactual ridership is calculated as $N_t^c = N_t/(1 + e^d \times \Delta p/p)$, where $\Delta p/p = -30\%$ is the discount rate. Entries to the right of the scheduling window (between 7:30 and 9 AM) are not affected by the EBD, so $N_t^c = N_t$.

Step 2. We fit N_t^c in the sample time window but outside the rescheduling window with a flexible smooth function of time t, which we proxy with a polynomial up to the 5^{th} order. The

fitted smooth curve represents the counterfactual entries, $\hat{N}_t^c$, over the entire sample window between 5:30 and 9 AM.

Step 3. The excessive bunching at time t during the bunching window (6:30-6:59 AM) is calculated as

$$\Delta\hat{N}_t^b = N_t - \hat{N}_t^c \times (1 + e^d \times \Delta p/p).$$

$\hat{N}_t^c \cdot e^d \cdot \Delta p/p$ accounts for the opt-in trips. The number of rescheduled trips at time t during the missing window (7-7:29 AM) is calculated as

$$\Delta\hat{N}_t^m = \hat{N}_t^c - N_t.$$

The bunching mass (B) and the missing mass (M) can be respectively calculated as

$$B = \sum_{t=6:30}^{6:59} \Delta\hat{N}_t^b, \quad M = \sum_{t=7}^{7:29} \Delta\hat{N}_t^m.$$

We verify whether B and M are sufficiently similar.

Step 4. The rescheduling elasticity in time t after 7 AM can be calculated as

$$\hat{e}_t^r = \frac{\Delta\hat{N}_t^m/\hat{N}_t^c}{-\Delta p/p}. \tag{8}$$

Confidence intervals are obtained by bootstrapping the entire process.

Panel C of Figure 8 plots the numbers of bunching and missing trips by five-minute bins. Missing trips are concentrated between 7 and 7:15 AM, while most of the bunching trips land between 6:35 and 6:59 AM. The graph shows that the chosen rescheduling window is sufficiently wide to capture most missing and bunching trips. The missing mass M is 995, and the bunching mass B is 1,263. Not only are B and M similar, but they are also rather small compared to the overall ridership. The total counterfactual ridership in the missing window is 27,430. The 30% EBD incentivized a mere 3.6% of the trips to reschedule in the missing window, or less than 1% during the morning rush hour (7-9 AM).

Panel D plots the rescheduling elasticity by five-minute bins (in green crosses) and non-parametric smooth fitted lines with bootstrapped 95% confidence intervals (in red lines). The rescheduling elasticity is close to 0.4 for trips originally planned for right after 7 AM but quickly drops to near zero by around 7:15 AM.

We carry out several robustness checks, which are described in Appendix B.7. First, we vary the bunching and missing windows. Second, we devise an approach that jointly estimates the demand and rescheduling elasticities. Third, we use the ridership in September 2015 as the coun-

terfactual. Results from all those checks are remarkably similar to those in the baseline.[20]

4 Welfare Impacts

4.1 Impacts on Consumer Welfare and Revenue

We evaluate the impacts of fare rise on consumer surplus, revenue, and congestion externality. Figure 9 Panel A presents a simplified illustration of the fare adjustment on consumer welfare depending on how users respond to the kinked budget constraint. Consumers demand subway trips S and a numeraire good C. We assume subway trips are homogeneous and have a unit price of p_0 under the original flat rate. The monthly budget constraint is represented by line $\bar{C}D_0$. A consumer chooses S_0 and obtains a utility level u_0.[21]

The budget constraint under the new fare structure with cumulative quantity discount is represented by $\bar{C}KD_{ra}$. The consumer faces a new listing price $p_L > p_0$ until her cumulative spending exceeds a pre-determined threshold, after which she qualifies for a discount rate δ. The consumer who chooses S_0 under the original price would choose S_{ra} and obtains a utility level u_{ra}, where $S_{ra} < S_0$ and $u_{ra} < u_0$.

An oblivious consumer ignores the discounts and thinks, mistakenly, that the price is always p_L, and she is on the budget line $\bar{C}D_{ob}$. Under this misperception, she thinks her optimal choice is at S'_{ob}, although she actually receives discounts, and her consumption is on the budget line at S_{ob}. Compared with rational consumers, oblivious consumers respond to a higher marginal price and take fewer subway trips ($S_{ob} < S_{ra}$), and incur a welfare loss because of that ($u_{ob} < u_{ra}$).

Like a rational consumer, an ironer can predict her monthly demand correctly and is aware of the quantity discount. However, she responds to the average price of her monthly expenditure, $\bar{p}$, instead of the marginal price. Graphically, the ironer chooses a point on the budget line at which the indifference curve is tangent with the linearized budget line with slope $-\bar{p}$. Her choice, denoted as S_{ir}, lies between the two polar cases of rational and oblivious, and so is her utility level u_{ir}.

Finally, a consumer is myopic if she responds to the current price she faces without taking the monthly budget into consideration. Her choices based on the *daily* budget are depicted in the lower-left corner of Panel A. At the start of the month, the price of a subway trip is p_L. Her

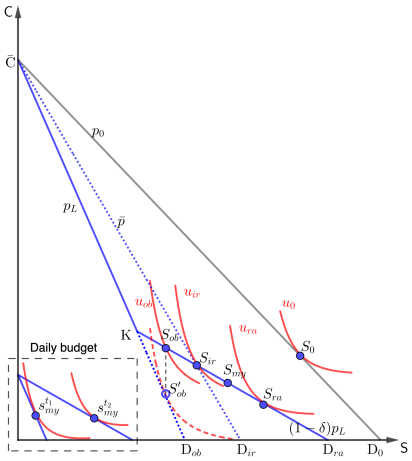
[20]Intuitively, the magnitude of the rescheduling elasticity is bounded from above by the size of the drop in ridership around the time cutoff. Figure 8 Panel A shows the drop is about 250 trips, while the counterfactual number of trips at 7 AM is around 750. For a rule-of-thumb calculation, we assume missing trips on the right and bunching trips on the left evenly divide up the gap. The upper bound of the rescheduling elasticity at 7 AM can be calculated as 250/2/750/0.3, which is around 0.56.

[21]We consider a consumer with a sufficiently large demand for subway trips such that she qualifies for some discount under the new fare structure.

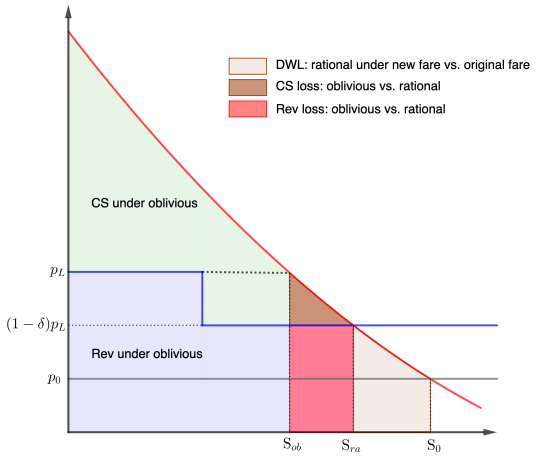
optimal choice is $s_{my}^{t_1}$. As the monthly proceeds and her cumulative expenditure on the subway exceeds the cutoff, she qualifies for the discounted price of $(1-\delta)p_L$, under which her optimal daily demand is $s_{my}^{t_2}$. The monthly total of the subway trips is S_{my}, which, as we show later, is close to S_{ir}.

Figure 9: Impacts of the Fare Structure Change: Illustration

Panel A: Demand for Subway Trips



Panel B: Welfare Implications



Note: Panel A illustrates the consumer's demand for subway trips under the original flat rate and the new fare structure and under various behavioral assumptions. Panel B illustrates the welfare impacts of the fare structure change, comparing the scenario in which the consumer responds rationally to the cumulative quantity discount and one in which the consumer is oblivious to the discounts. See text for detailed descriptions.

Panel B illustrates the welfare impacts of the fare structure change under two polar behavioral types: rational and oblivious. For simplicity, we assume the marginal cost of a subway trip is zero and do not consider the negative externality due to crowding. The area in pink illustrates the increased deadweight loss under the new fare structure with rational consumers. The increase in deadweight loss is a result of fewer subway trips. With zero social marginal cost, those trips are welfare-improving at the societal level.

If consumers are oblivious, the welfare loss is larger. The area in brown represents the additional deadweight loss in consumer welfare compared with the rational case, while the area in red represents lost revenue. The area in green represents the consumer surplus, and revenue is colored in purple.

Panel B also shows that it is intuitive to calculate welfare impacts. Price and discount schedules are known. Demand for subway trips and relevant prices under various behavioral assumptions can be backed out by the estimated demand elasticity.

We use the ridership in September 2014 as the pre-fare-adjustment baseline. The data in

September 2014 covers only one week. We populate the existing data to cover the entire month.[22] The simulated pre-adjustment monthly ridership covers 6.7 million users. An average user has 18.4 trips covering a distance of 289 km during the month (Table 5 Column 1). For each user, we sort trips chronically and calculate the corresponding discount rates.

We classify users by their travel patterns using the same set of predictors as we did for those in April 2015. The $K-$means clustering algorithm yields the same set of card types and similar distributions. Appendix Table C.1 summarizes the characteristics of users by different types.

Figure 3 shows that demand elasticity differs by user type but not by trip type within each user type. Therefore, we assign each card type with one demand elasticity. We assign the same set of rescheduling elasticity, estimated in Section 3.4, to all users. We impose a constant-elasticity demand function and calculate the counterfactual number of trips under the new fare structure and under different behavioral responses to cumulative quantity discounts.

Columns 2 and 3 of Table 5 report the impacts of the fare structure change on ridership and welfare for the average user, assuming all users respond only to the listing price (oblivious). The average listing price increases from 2 yuan to about 4.7 yuan per trip. Consequently, the average monthly number of subway trips declines by 22%, from 18.4 to 14.3. The average passenger-kilometer decreases by a larger 25% because longer trips experience a larger percentage increase in fare. The average user (not the average trip) qualifies for an average discount rate of 3.3%. Average out-of-pocket expenditure, which is equal to the average revenue per user, increases by 56% to 57 yuan. The revenue would increase by 1.6 billion yuan per year, or equivalent to 10% of the annual operating cost of the system. Compared with the original 2-yuan flat rate, the average consumer welfare declines by 34 yuan. Given revenue increases by 21 yuan, the deadweight loss (the decrease in consumer welfare minus the increase in revenue) is 13 yuan per user per month.

The remaining columns show the impacts on revenue and consumer welfare if users respond to the discounts in some way. If consumers are rational, the average consumer takes 16.4 trips per month under the new fare structure, or 15% more than the average oblivious consumer. Revenue would be 9% higher (63 yuan per user per month versus 57 yuan), and the deadweight loss would only be 42% of that under oblivious (5.4 versus 13). The choices and welfare consequences if users are myopic or ironing are similar,[23] and lie between the cases with users assumed to be oblivious or rational.

[22] We need full-month data to calculate the discount rate and how users respond to the nonlinear monthly budget. Appendix D.1 documents how we simulate the data.

[23] Results are similar under ironing and myopic because the average of instantaneous marginal prices (to which myopic users respond) is similar to the month-end average price (to which ironers respond). They are not identical because subway fare is a step function of distance and because of curvatures in the demand function.

Table 5: Aggregate Impacts of the Fare Structure Change

		new fare schedule				
	orig.	assuming	% chg	alternative behavioral responses		
	flat rate	oblivious	from orig.	rational	ironing	myopic
	(1)	(2)	(3)	(4)	(5)	(6)
Ridership (per user per month)						
# of trips	18.4	14.3	-22.3%	16.4	15.1	15.1
Total distance (km)	289	216	-25.3%	249	229	231
Expenditure (yuan)						
before discount	36.8	66.7	81.3%	76.6	70.4	70.8
after discount	36.8	57.3	55.7%	62.5	59.3	59.5
Discount rate	-	3.3%	-	4.3%	3.6%	3.6%
Welfare impacts (yuan per user per month)						
Δ Consumer surplus	-	-33.5	-	-31.1	-32	-32.4
Δ Revenue	-	20.5	-	25.7	22.5	22.7
Δ DWL: -(ΔRev+ΔCS)	-	13	-	5.4	9.5	9.7
Δ Congestion externality	-	8.4	-	4.6	7.0	6.9

Note: The table summarizes the aggregate welfare impacts of changing the fare structure from a 2-yuan flat rate to the current pricing schedule. Pre-fare-adjustment monthly ridership patterns are based on data from September 2014. There are 6.7 million unique users. The table reports consumer-level monthly averages.

4.2 Congestion Externality

Regardless of how consumers respond to discounts, Table 5 shows that the fare structure change led to substantial reductions in subway trips (ranging from 11% if consumers are rational to 22% if oblivious). At least some of those trips will be fulfilled by other modes of transportation. Most alternative modes would involve using surface roads, leading to increased road congestion. In this subsection, we quantify the magnitude of congestion externality using our estimates as well as parameters we borrow from other studies.

We start with the marginal cost of traffic congestion (MECC) due to an additional one vehicle-kilometer. Following Yang et al. (2020), the MECC at time t can be written as

$$MECC_t = o \cdot VOT \cdot T_t \cdot \frac{\varepsilon_t}{1-\varepsilon_t}, \tag{9}$$

where o is the average number of passengers in a vehicle (vehicle occupancy), which is 1.32 according to the 2014 Beijing Household Travel Survey. VOT is the value of time measured in yuan per hour. Assuming 50% of market wage in Beijing,[24] Gu et al. (2021b) estimate average hourly wage in Beijing around that time was 46.2 yuan, so VOT is 23.1 yuan.

T_t is the inverse of speed, measured in hours per kilometer. We use the hourly speed data

[24] Assuming the value of time in travel to be 50% of hourly wage is a rule of thumb in the literature (e.g.,Small and Verhoef, 2007; Parry and Small, 2009), although recent estimates find a larger value of time (e.g., Kreindler, 2020).

from Yang et al. (2020), which are from a set of road monitoring stations on Beijing's highways and ring roads in 2014.[25] $\varepsilon_t = -(\partial Speed/\partial Density) \cdot (Density_t/Speed_t)$ is the elasticity of speed with regard to density. Yang et al. (2020) estimate $\partial Speed/\partial Density$ to be -1.136. Road density and speed at time t is the weighted average of monitor-station-level data, where the weight is the share of subway trips in the corresponding geographical area of Beijing.[26]

The final piece of information we need is the fractions of reduced subway trips converted to other modes of transportation. Those conversion rates are essentially tied to the substitutability among different modes. Without a good source for those parameters, we assume 50% of the missing subway trips are diverted to bus trips, 25% to car trips, and the remaining 25% to various two-wheel vehicles, or no trip at all. Trips in the last category are assumed to have no impact on road congestion.

The last row of Table 5 reports the increases in congestion externality, denoted in yuan per user per month, as the fare is changed from the 2-yuan flat rate to the current structure. In the baseline scenario where consumers are oblivious to quantity discounts, the additional congestion externality amounts to 8.4 yuan per user per month. This is around the same magnitude as the deadweight loss (13 yuan per user per month). Yang et al. (2020) estimate that if pollution, green gas emission, and accidents are considered, the total negative externality is 2.7 times larger than the congestion externality alone. Therefore, negative externality due to increased road traffic is an important component of the welfare impact. The negative externality will be smaller if consumers at least partly respond to the discounts, as the decline in subway ridership will be milder. For example, if consumers are fully rational, the additional congestion externality will be 4.6 yuan per user per month, which is 45% smaller than that under users being oblivious.

4.3 Distributional Impacts

Table 6 reports the distributional impacts of the fare structure change on different types of users. To keep the table concise, we consider only two polar behavioral responses to the quantity discounts: oblivious and rational.

Infrequent users and those who mostly travel on weekends or during non-rush hours usually do not qualify for any discount, so alternative behavioral models make little difference in the impacts on ridership and welfare. The demand elasticity for these groups is also relatively small, so there are relatively small changes in ridership, substantial increases in revenue, and relatively

[25] Several adjustments are made to the speed data in Yang et al. (2020) to fit our needs. See Appendix D.3 for details.

[26] Beijing has five ring roads, which divide Beijing into six regions according to their distance to the city center. We group traffic monitor stations by those six regions and weigh each group by the aggregate length of subway trips that take place within the region. A subway trip may extend multiple regions; the trip's distance is proportionated to each of the regions it crosses.

small efficiency loss and congestion externality.

On the other hand, rush-hour commuters substantially slash their subway trips. The average user in this category reduces her number of subway trips by 34% if they are oblivious to the quantity discounts. The large decline is due to two main reasons. First, the demand elasticity for this group is particularly large, as is shown in Figure 3. Second, as frequent users, they qualify for a substantive amount of discounts. Ignoring those discounts results in a substantial departure from the optimal choices and large welfare losses. For the average user in a month, revenue increases by 30 yuan, but this is at the cost of reducing consumer surplus by 75 yuan and a congestion externality of 30 yuan. In this case, consumers' behavioral responses to discounts matter significantly. Consumers would take advantage of the discount and cut trips less if they were rational. As a result, the deadweight loss would be 63% smaller (16.5 vs. 44.9), and the additional congestion externality would be 49% less (15.1 v. 29.5).

Similar patterns are observed for frequent but less regular commuters. The deadweight loss would have been 79% smaller if they were rational instead of oblivious. Compared with regular commuters, this group of users has a smaller demand elasticity, so price increases lead to smaller changes in ridership. Their subway trips are also less concentrated during weekday peak hours, so trips that were diverted to other modes cause smaller increases in road congestion.

Table 6: Distributional Impacts of the Fare Structure Change

	Infrequent users			*Weekenders*			*Weekday non-rushers*		
	orig.	new fare schedule		orig.	new fare schedule		orig.	new fare schedule	
	rate	oblivious	rational	rate	oblivious	rational	rate	oblivious	rational
# of trips	2.6	2.2	2.2	13.1	10.5	10.8	17.4	14.7	15.4
Exp. or Rev. (yuan)	5.3	10.5	10.5	26.2	47.3	48.1	34.8	63.9	65.9
ΔCS (yuan)	-	-6.8	-6.8	-	-29.9	-29.6	-	-37.8	-37
ΔDWL (yuan)	-	1.6	1.6	-	8.8	7.7	-	8.7	5.9
ΔCong. Exter. (yuan)	-	0.9	0.9	-	5.1	4.7	-	5.5	4.2

	Rush-hour commuters			*Less-regular commuters*		
	orig.	new fare schedule		orig.	new fare schedule	
	rate	oblivious	rational	rate	oblivious	rational
# of trips	41.2	27.0	34.5	54.9	44.6	51.6
Exp. or Rev. (yuan)	82.5	112.1	131.5	109.9	162.6	179.6
ΔCS (yuan)	-	-74.5	-65.5	-	-84.2	-76.2
ΔDWL (yuan)	-	44.9	16.5	-	31.5	6.5
ΔCong. Exter. (yuan)	-	29.5	15.1	-	21.3	8.7

Note: The table summarizes the distributional welfare impacts of changing the fare structure from a 2-yuan flat rate to the current pricing schedule for different types of users. Pre-fare-adjustment monthly ridership patterns are based on data from September 2014. User types are determined by ridership patterns using a $K-$means clustering algorithm.

4.4 Discussion of the Welfare Impacts

Overall, calculations in this section show that due to inelastic demand, the fare rise effectively raised revenue at a relatively small cost of the ridership. Consequently, the incidence is largely shouldered by consumers. Consumer welfare loss is large relative to the increase in revenue, resulting in a large deadweight loss. In addition, reduced subway ridership leads to more vehicles on surface roads, and the resulting increase in congestion externality is at the same magnitude as the deadweight loss. Furthermore, the small rescheduling elasticity indicates that a higher price for peak-hour trips does little to incentivize trips to reschedule to less busy hours. The impacts are not evenly distributed across different users. Less frequent users spent a lot more in percentage terms, while frequent users experienced larger increases in expenditure in dollar terms and are the source of large deadweight losses.

A few aspects are not considered in the welfare calculation. We assume that providing an additional subway trip has zero marginal cost, which essentially assumes the system has no crowding externality. Using the same data, Gu and Zou (2023) shows that the elasticity of time cost of a subway trip with regard to the number of passengers is rather small even during rush hours, although they do not estimate utility loss due to discomfort in crowded platforms and crammed passenger cars. We also do not calculate other externalities besides road congestion. Yang et al. (2020) estimate that if pollution, green gas emission, and accidents are considered, the total negative externality is 2.7 times larger than the congestion externality alone. Finally, we do not consider the inefficiency created by the distortive taxation that subsidizes the subway system.

How consumers respond to the quantity discount greatly affects aggregate and distributional impacts. Being oblivious to the discounts makes consumers cut more subway trips, which leads to lower revenue, larger consumer welfare loss, and a larger congestion externality. The welfare cost of ignoring quantity discounts is particularly high for frequent users.

When consumers do not respond optimally to discounts, the current design of the fare structure may not achieve the desired outcomes. Next, we consider whether alternative fare structures could achieve higher aggregate and allocative efficiency.

5 Alternative Fare Structures

5.1 Alternative Fare Structures under Consideration

We evaluate ridership and welfare under two alternative fare structures. The first is an alternative flat rate without the quantity or early bird discount. It is the simplest possible fare structure that eliminates any concern for behavioral responses to complex pricing schedules. Second, we

replace the EBD with a peak-hour premium,[27] while keeping other aspects of the current fare structure. Peak-hour premium is popular among transit systems and is often justified by its role in reducing system crowdedness during rush hours, either by diverting subway trips to other modes or moving them to less busy hours. While the former channel is governed by the demand elasticity, the latter is governed by the rescheduling elasticity, which we estimate in Section 3.4.[28]

For both alternative fare structures, we calculate fare levels that would generate the same level of revenue as under the current fare structure, specifically for each behavioral type in response to the nonlinear monthly budget. A two-layered iterative algorithm is used to determine the fare level. Starting with an initial guess of the fare level under the alternative fare structure, we calculate each user's demand for subway trips that is consistent with the price she perceives according to the specific behavioral type. We then calculate the aggregate revenue and update the fare level until the resulting revenue is the same as that under the current fare structure. Appendix Section D.2 describes the details of the algorithm.

Table 7 Panel A reports fare levels under alternative fare structures. In the interest of space, the table reports two polar behavioral types: rational and oblivious. Fares levels, associated ridership, and welfare with consumers being myopic or ironing are reported in Appendix Table D.1.

If users are oblivious, the current fare structure generates 57.3 yuan per user per month. The alternative flat rate needs to be set at 3.78 per trip, while the fare structure with peak premium needs to be set at 4.52 yuan as the listing price for a 6-km trip during peak hours.[29] To make fare levels comparable across alternative pricing schedules, the last row in Panel A reports the listing price per kilometer under the pre-adjustment ridership pattern. The unit price is the lowest under the alternative flat rate, partly because no discount is offered on this price. Columns 5 through 7 report required fare levels if users are rational.

5.2 Ridership and Welfare under Alternative Fare Structures

Panel B reports ridership under alternative fare structures, and Panel C reports associated welfare changes relative to the original 2-yuan flat rate. All numbers reported in these two panels are per user per month.

If users do not respond to the quantity discounts, the average user has 14.3 trips and spends 57.3 yuan under the current fare structure. Compared with the 2-yuan flat rate, While the revenue

[27] Peak hours are between 7 and 9 AM and between 5 and 7 PM on weekends. Fares during peak hours are set to be twice as much as off-peak hours.

[28] Admittedly, many other common components in transit fare pricing are not considered here. Evaluating such components' roles would require parameters not estimated in the paper. For example, many transit systems offer a monthly pass. However, findings in this paper provide little insight into how users may choose a monthly pass and decide their ridership accordingly if such an option is offered.

[29] The fare for the same trip is 2.26 yuan during off-peak hours. A 6-km subway ride has a listing price of 3 yuan under the current fare structure.

Table 7: Aggregate Ridership and Welfare under Alternative Fare Structures

			Oblivious			Rational	
	orig. flat rate	current fare	alt. flat rate	peak/ off-peak	current fare	alt. flat rate	peak/ off-peak
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Alternative prices (yuan)							
Flat rate	2		3.78			4.29	
Listing p for a 6 km ride		3		4.52 (peak)	3		4.46 (peak)
Avg. listing p/km	0.13	0.30	0.24	0.32	0.30	0.27	0.31
Panel B: Ridership (per user per month)							
Monthly revenue (yuan)	36.8	57.3	57.3	57.3	62.5	62.5	62.5
# of trips	18.4	14.3	15.2	14.3	16.4	14.6	16.4
Total distance (km)	289	216	239	217	249	229	250
Avg. discount rate (%)	-	3.3	-	3.5	4.3	-	4.5
Panel C: Change in welfare compared with original flat rate (yuan per user per month)							
Revenue increase	-	20.5	20.5	20.5	25.7	25.7	25.7
Consumer welfare loss	-	33.5	30.0	36.3	31.1	37.7	35.9
Deadweight loss	-	13.0	9.4	15.7	5.4	12.0	10.2
Congestion externality	-	8.4	6.0	13.7	4.6	7.0	9.9

Note: This table summarizes aggregate ridership and welfare under alternative fare structures, assuming consumers are oblivious or rational to the quantity discounts. Pre-fare-adjustment monthly ridership patterns are based on data from September 2014. Deadweight loss is the consumer's utility loss minus the operator's gain in revenue. See Appendix Table D.1 for aggregate ridership and welfare impacts with myopic and ironing users.

increases by 20.5 yuan, consumer welfare declines by 33.5 yuan, resulting in a deadweight loss of 13 yuan. In addition, some subway trips are diverted to road traffic, which generates a congestion externality of 8.4 yuan.

In this case, a revenue-preserving flat rate performs better in the aggregate. Users take more trips (15.2 v. 14.3). Consumer welfare loss and the deadweight loss are, respectively, 10% and 28% smaller. With fewer trips diverted to surface roads, the welfare loss due to increased road congestion is 29% smaller. The flat rate outperforms the current fare structure because oblivion to discounts disproportionately hurts commuters, who derive high values from subway trips and are more likely to travel during peak hours, when a diverted subway trip causes a larger negative externality on surface roads.

However, if users are rational, the current fare structure outperforms the iso-revenue flat rate. Compared with the current fare structure, the average rational user takes 11% fewer trips under the flat rate. Consumer welfare loss, deadweight loss, and the increase in congestion externality is, 21%, 122%, and 52% higher, respectively, under the flat rate. The current fare structure is designed to cross-subsidize regular, frequent commuters who mostly travel during peak hours. The simulations here show that it works in theory, although in practice, it performs worse as

users do not rationally respond to the incentives.[30]

In both cases, adding a peak premium to the current fare structure results in larger welfare losses and negative congestion externalities (comparing Columns 2 and 4, and Columns 5 and 7). This is because peak/off-peak pricing further penalizes frequent users, who have high-valued trips during rush hours. Although, as shown in Appendix Figure D.1, it is the most effective in reducing peak-hour system load, which it achieves by pushing peak-hour subway trips to other transportation modes (or no travel at all) rather than diverting them to a less busy time.

6 Conclusion

The design of transit fares serves multiple policy goals and is often complex in nature. This paper estimates consumer responses to a substantial fare structure adjustment in Beijing's subway. We find that the demand elasticity for subway trips is small, with larger demand elasticity found among regular commuters, schedules for peak-hour trips are inflexible, and early-bird discounts have a negligible effect on diverting trips to non-peak hours.

Consumers do not seem to respond to the cumulative quantity discounts. We can soundly reject that consumers are forward-looking and optimize on the monthly budget. Evidence also does not support consumers adopting heuristic decision-making models by responding to the average or instantaneous marginal prices. Estimating a statistical mixture model indicates that consumers overwhelmingly disregard discounts and respond only to the listing price.

The empirical estimates are then used to quantify the aggregate and distributional impacts of the fare structure change. Inelastic demand indicates a substantial transfer from consumer surplus to the operator's revenue at a relatively small cost to the total ridership. Due to consumers' unresponsiveness, quantity discounts are ineffective in cross-subsidizing frequent users, translating into a large social welfare loss. We show the welfare impacts of the new fare structure would differ substantially under alternative consumer behavior.

The paper's empirical findings provide key elements to the sensible design of fare structures. The current fare structure could achieve high social welfare under the ideal scenario where consumers respond rationally to all the embedded incentives. However, a revenue-preserving flat rate, whose simplicity eliminates confusion and optimization frictions from a complex fare structure, would be preferred if consumers are less than fully rational. Finally, because the rescheduling elasticity is small, a peak-hour premium does little to divert trips to less busy hours and generates large negative externalities on surface road congestion.

[30] Appendix Table D.1 shows that the current fare structure and the corresponding iso-revenue flat rate achieve similar levels of efficiency when users either respond to the month-end average price (ironing) or the instantaneous marginal price (myopic). Appendix Table D.2 reports the distributional impacts of alternative fare structures on different user types.

References

Anderson, M. L. (2014): "Subways, strikes, and slowdowns: The impacts of public transit on traffic congestion," *American Economic Review*, 104, 2763–96.

Balboni, C., G. Bryan, M. Morten, and B. Siddiqi (2020): "Transportation, gentrification, and urban mobility: The inequality effects of place-based policies," *Preliminary Draft*, 3.

Beijing Transport Institute (2015): *Household Travel Surveys*, Beijing Transport Institute.

Borenstein, S. (2009): "To what electricity price do consumers respond? Residential demand elasticity under increasing-block pricing," *Preliminary Draft April*, 30, 95.

——— (2012): "The redistributional impact of nonlinear electricity pricing," *American Economic Journal: Economic Policy*, 4, 56–90.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014): "Robust nonparametric confidence intervals for regression-discontinuity designs," *Econometrica*, 82, 2295–2326.

Cattaneo, M. D., N. Idrobo, and R. Titiunik (2019): *A practical introduction to regression discontinuity designs: Foundations*, Cambridge University Press.

Cervero, R. (1990): "Transit pricing research: A review and synthesis," *Transportation*, 17, 117–139.

Chen, T., Y. Gu, and B. Zou (2022): "Delineating China's Metropolitan Areas Using Commuting Flow Data," *Available at SSRN*.

Chen, Y. and A. Whalley (2012): "Green infrastructure: The effects of urban rail transit on air quality," *American Economic Journal: Economic Policy*, 4, 58–97.

Chetty, R., A. Looney, and K. Kroft (2009): "Salience and taxation: Theory and evidence," *American economic review*, 99, 1145–77.

Davis, L. W. (2021): "Estimating the price elasticity of demand for subways: Evidence from Mexico," *Regional Science and Urban Economics*, 87, 103651.

Gendron-Carrier, N., M. Gonzalez-Navarro, S. Polloni, and M. A. Turner (2022): "Subways and urban air pollution," *American economic journal: Applied economics*, 14, 164–96.

Gu, Y., N. Guo, J. Wu, and B. Zou (2021a): "Home Location Choices and the Gender Commute Gap," *Journal of Human Resources*, 1020–11263R2.

Gu, Y., C. Jiang, J. Zhang, and B. Zou (2021b): "Subways and road congestion," *American Economic Journal: Applied Economics*, 13, 83–115.

Gu, Y. and B. Zou (2023): "Congestion and crowding externalities of public transit: Evidence from Beijing." .

Heblich, S., S. J. Redding, and D. M. Sturm (2020): "The making of the modern metropolis: evidence from London," *The Quarterly Journal of Economics*, 135, 2059–2133.

Ito, K. (2014): "Do consumers respond to marginal or average price? Evidence from nonlinear

electricity pricing," *American Economic Review*, 104, 537–63.

Ito, K. and S. Zhang (2020): "Reforming inefficient energy pricing: Evidence from china," *NBER Working Paper*.

Jessoe, K. and D. Rapson (2014): "Knowledge is (less) power: Experimental evidence from residential energy use," *American Economic Review*, 104, 1417–38.

Kleven, H. J. (2016): "Bunching," *Annual Review of Economics*, 8, 435–464.

Kreindler, G. (2020): "Peak-hour road congestion pricing: Experimental evidence and equilibrium implications," *Unpublished paper*.

Li, S., Y. Liu, A.-O. Purevjav, and L. Yang (2019): "Does subway expansion improve air quality?" *Journal of Environmental Economics and Management*, 96, 213–235.

Liebman, J. B. (1998): "The impact of the earned income tax credit on incentives and income distribution," *Tax policy and the economy*, 12, 83–119.

Liebman, J. B. and R. J. Zeckhauser (2004): "Schmeduling," Working paper.

Lu, Y., X. Shi, J. Sivadasan, and Z. Xu (2021): "How Does Improvement in Commuting Affect Employees? Evidence from a Natural Experiment," *Review of Economics and Statistics*, 1–47.

Ma, Z., H. N. Koutsopoulos, T. Liu, and A. A. Basu (2020): "Behavioral response to promotion-based public transport demand management: Longitudinal analysis and implications for optimal promotion design," *Transportation Research Part A: Policy and Practice*, 141, 356–372.

Parry, I. W. and K. A. Small (2009): "Should urban transit subsidies be reduced?" *American Economic Review*, 99, 700–724.

Pelletier, M.-P., M. Trépanier, and C. Morency (2011): "Smart card data use in public transit: A literature review," *Transportation Research Part C: Emerging Technologies*, 19, 557–568.

Reiss, P. C. and M. W. White (2005): "Household electricity demand, revisited," *The Review of Economic Studies*, 72, 853–883.

Saez, E. (2010): "Do taxpayers bunch at kink points?" *American economic Journal: economic policy*, 2, 180–212.

Sexton, S. (2015): "Automatic bill payment and salience effects: Evidence from electricity consumption," *Review of Economics and Statistics*, 97, 229–241.

Small, K. A. and E. T. Verhoef (2007): *The economics of urban transportation*, Routledge.

Tsivanidis, N. (2019): "Evaluating the impact of urban transit infrastructure: Evidence from bogota's transmilenio," Tech. rep., UC Berkeley (mimeo), 2020.[Google Scholar].

Vickrey, W. (1980): "Optimal transit subsidy policy," *Transportation*, 9, 389–409.

Vickrey, W. S. (1963): "Pricing in urban and suburban transport," *The American Economic Review*, 53, 452–465.

World Bank (2009): *World development report 2009: Reshaping economic geography*, The World Bank.

Yang, J., A.-O. Purevjav, and S. Li (2020): "The marginal cost of traffic congestion and road pricing: evidence from a natural experiment in Beijing," *American Economic Journal: Economic Policy*, 12, 418–53.

Zárate, R. D. (2022): *Spatial misallocation, informality, and transit improvements: Evidence from mexico city*, The World Bank.

Online Appendix

A Data Summary

Appendix Table A.1 describes the subway ridership data used in this paper. We mainly use three waves of data that cover a period of two years. Each wave covers the universe of all trips that use a smartcard to settle the fare during the corresponding time frame. The three waves of data include two full weeks and one full month. The week between September 15 and 21, 2014, is the only wave that was before the fare structure adjustment. The baseline analysis focuses on changes in ridership between the week of September 15, 2014, and the month of April 2015. The full-month ridership data from April 2015 allows us to analyze how consumers respond to the non-linear monthly budget. Ridership data in the week of September 12, 2016, are used to estimate the rescheduling elasticity because the early-bird discount was not implemented until December 2015.

Table A.1: Data Description and Summary

Panel A: Data used in the paper					
period	dates	# of workdays	ridership (mil./day)	# of non-workdays	ridership (mil./day)
September 2014	9/15-9/21	5	4.5	2	3.3
April 2015	4/1-4/30	21	4.6	9[1]	3.0
September 2016	9/12-9/18	4	5.3	3[2]	3.1

Panel B: Summary statistics of subway ridership in September 2014				
	workday		non-workday	
	mean	median	mean	median
distance (km)	15.96	14.38	16.87	14.97
time cost (minutes)	39.27	37.14	43.17	39.32
speed (km/h)	23.94	24.35	22.78	23.31

Notes: [1] 4/6 (Monday) is Qingming Holiday. [2] 9/15-9/17/2016 is Mid-Autumn Holiday. 9/18/2016 (Sunday) is workday.

There are several national holidays in our sample, during which the ridership pattern could be different. When a national holiday occurs, workdays and weekends are reorganized to make non-workdays contiguous. Such reorganization will sometimes make a Saturday or a Sunday into a workday. For example, Mid-autumn Day in 2016 landed on September 15, which was a Thursday. The Mid-autumn holiday was extended to three days to include the following weekend. In order to make the non-work days contiguous, Friday (September 16) was switched with Sunday (September 18), and the latter became a workday. However, firms can decide whether to extend the holiday to include that Sunday, and many do. In our analyses, we flag all those affected days as

potentially affected by the holiday. Regarding Mid-autumn day in 2016, we treat all days between September 15 and September 18 as nonwork days.

Panel A of Table A.1 reports the daily ridership separately for workdays and non-workdays. The average ridership was between 4.5 million and 5.3 million on a workday and around 3 million on a non-workday. Despite the substantial fare rise, subway ridership has been increasing over time. Panel B reports the mean and median distance, time cost, and speed of the trips. The median subway trip during a workday has a distance of about 14 km and takes about 37 minutes between tapping into the origin station and tapping out of the destination station. This yields a median speed of 24 km per hour. Trips on non-workdays are slightly longer in distance. Probably due to less frequent services, the average speed is slightly lower.

In addition, data from the week of September 14, 2015, are used to conduct a few robustness checks. First, they are used to conduct robustness checks for the OD-pair regression discontinuity estimations of the demand elasticity. Most of the results from those robustness checks are presented in Appendix Figure B.2. Second, they are used to serve as the control group in an alternative specification to estimate the rescheduling elasticity. Results from those estimations are presented in Appendix Figure B.7.

B Additional Robustness Checks

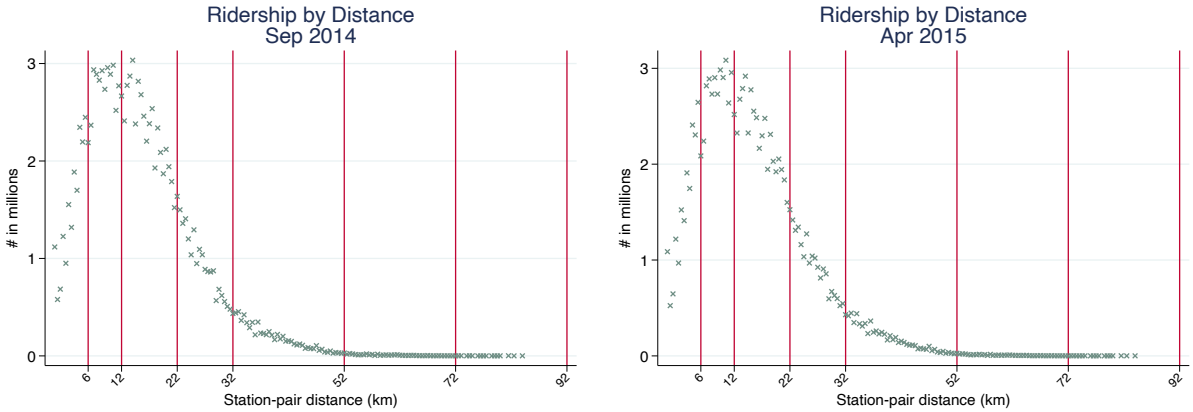
B.1 Robustness Checks to Station-pair Regression Discontinuity Estimations of the Demand Elasticity

Ridership by station-pair distance bins. Figure 1 shows the log *changes* in ridership between September 2014 and April 2015 in station pairs by 500-meter distance bins. There are visually sharp discontinuities around distance cutoffs. Appendix Figure B.1 plots ridership *levels* by station pairs in 500-meter bins, separately for September 2014 (Panel A, weekly ridership converted into monthly ridership by multiplying 30/7) and April 2015 (Panel B).

Ridership as a function of distance is a smooth curve in September 2014. There is some visual evidence of discontinuity around distance cutoffs in April 2015, but those discontinuities are less sharp than those in Figure 1. There is substantial heterogeneity in ridership across station pairs. Some station pairs are popular and have a large ridership. Such heterogeneity blurs the discontinuity in the number of trips. Taking the difference in ridership within each station pair eliminates cross-sectional heterogeneity and generates sharp discontinuities. Therefore, our baseline model constitutes a regression discontinuity design in difference.

Robust checks to station-pair RD estimations. In the baseline, we estimate Equation 1 where the running variable ($dist_{od}$) is fitted with a flexible *global* polynomial function $f(.)$. This ap-

Figure B.1: Station-pair Ridership by Distance Bins



Note: Each dot represents weekly total ridership in origin-destination pairs within a 500-meter bin. The left panel shows the ridership of the week in September 2014, the right panel shows that of the week in April 2015. Red vertical lines indicate the distance thresholds for a higher fare.

proach has several advantages. First, it is simple and transparent. Second, it allows for conveniently combining multiple cutoffs in a single regression, in which the coefficient associated with the log price is the demand elasticity that we are interested in estimating.

Calonico et al. (2014) propose an RD estimation where the running variable is fitted by nonparametric local linear regressions, which they show have good statistical properties. We re-estimate Equation 1 using this approach at each cutoff as well as with all cutoffs combined.[1] Appendix Table B.1 reports the results. The demand elasticity estimates are remarkably similar to those from global polynomials, reported in Table 2.

Table B.1: Demand Elasticity Using Local Linear Regression Discontinuity Design

	(1)	(2)	(3)	(4)	(5)
		at distance cutoff			at all cutoffs
	6 km	12 km	22 km	32 km	all
e^{RD}	-0.472	-0.442	-0.303	-0.406	-0.398
	(0.122)	(0.093)	(0.099)	(0.191)	(0.073)
Sample range (km)	[-3,3]	[-3,3]	[-5,5]	[-5,5]	-

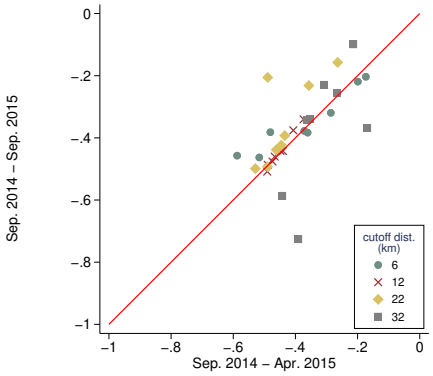
Note: The table reports results from estimating Equation 1 using regression discontinuity models with local linear regression, triangular kernel, and optimal bandwidth. Robust standard errors are in parentheses. The dependent variable is the log change in ridership in the same OD pair between September 2014 and April 2015. All regressions are weighted by the ridership in September 2014.

We also run a series of robustness checks of OD-pair RD regressions based on the baseline specification in Table 2. Separately for each distance cutoff, we vary the degree of polynomials

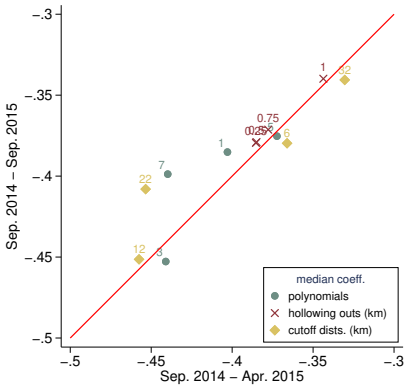
[1] In the `rdrobust` package, demand elasticity can be directly estimated by using a fuzzy RD design, where the log price of the station pair is instrumented by an indicator for the side of the cutoff the station pair is located.

Figure B.2: Summary of OD-pair RD Robustness Checks

Panel A: Estimated elasticity



Panel B: Median elasticity by group



Note: Estimated elasticity using Equation 1 at four cutoff distances, separately for two periods: (1) from September 2014 to April 2015, (2) from September 2014 to September 2015. Panel A shows all eight versions of the estimate for each cutoff that vary by degrees of polynomials and the size of the donut hole. Panel B shows median estimated elasticity by estimation group.

(up to 1^{st}, 3^{rd}, 5^{th}, and 7^{th}, respectively), and the size of the hollowing out region around the cutoff (0.25 km, 0.5 km, 0.75 km, and 1 km in radius, respectively). We also use data from the week of September 15, 2015, as a robustness check.

Appendix Figure B.2 summarizes the results from those robustness checks. Panel A reports the results of all 64 estimations. Most estimates are tightly distributed except for a few estimations from the 32-km cutoff. Estimates around the 32-km cutoff are noisy, as can be seen from Figure 1, because the number of trips around the cutoff is relatively small. Panel B reports the median estimates in each series (by cutoff distance, degree of polynomials, or size of the donut hole). All estimates are closely bounded between -0.3 and -0.45. In both panels, estimates from the two periods are closely distributed along the 45-degree line, which means that the elasticity is stable over this relatively short period.

B.2 Estimating the Bunching Elasticity

One concern with the OD pair RD design is that passengers may shorten their trips and bunch below the distance cutoff to reduce cost. Table B.2 illustrates how the presence of strategic bunching biases the RD estimator. Consider two OD pairs with similar distances but lie on different sides of the distance cutoff, denoted as OD_L and OD_R. In period $t = 0$, all OD pairs have a flat fare p. For simplicity, assume both OD pairs initially have the same number of trips Q. In period $t = 1$, the fare of OD_R doubles to $2p$ while that of OD_L remains unchanged. Due to the higher fare in OD_R, x trips are not taken, and y trips bunch to OD_L. Bunching reduces trips in OD_R and increases

trips in OD_L, but reflects little change in the actual use of the subway as measured by passenger mileage. Let e, e^{RD}, and e^b denote, respectively, the demand elasticity without including strategic bunching, the elasticity estimated using the OD pair RD design, and the bunching elasticity, which measures the percent of trips in OD_R bunched to OD_L in exchange for a one-percent saving on fare. The example in Table B.2 shows that $e = \frac{-x}{Q}$, $e^b = -\frac{y}{Q}$, and $e^{RD} = \frac{-x-y}{Q} - \frac{y}{Q} = \frac{-x-2y}{Q}$. Therefore, e^{RD} over-estimates e by $2e^b$.

Table B.2: Bunching and Demand Elasticity: An Example

	$t=0$		$t=1$	
	OD_L	OD_R	OD_L	OD_R
price	p	p	p	$2p$
# of rides	Q	Q	$Q+y$	$Q-x-y$

Note: The table illustrates how bunching affects the estimation of demand elasticity. OD_L and OD_R are two OD pairs with similar distances but lie on different sides of the distance cutoff. In $t = 0$, both OD pairs are priced at p and both have Q trips; in $t = 1$, the fare of OD_R increases to $2p$. x indicates the reduction in trips in OD_R due to the fare rise. y indicates the number of trips in OD_R that are bunched to OD_L.

Leveraging the panel structure of subway trips in the week of September 15, 2014, and the full month of April 2015, we present a way to directly estimate e^b and thus recover the demand elasticity e as $e = e^{RD} - 2e^b$.[2] At each cutoff, we estimate the following equation:

$$\frac{\Delta N_i^l}{N_{i14}^r} = e^b \cdot T_i \cdot \frac{p_r - p_l}{p_l} + \varepsilon_{it}. \tag{B.1}$$

T_i is a binary variable indicating whether a user i belongs to the treated group ($T_i = 1$) or the control group ($T_i = 0$). Specifically, for each cutoff c, the treated group is defined as those who have the majority of their trips in 2014 in a pre-specified "treated window" $[c, c+w]$. The treated window includes OD pairs to the right of but within w km from the cutoff. The corresponding "bunching window" is defined as $[c-w, c)$, which is intended to catch shortened trips that land to the left of the cutoff. The treated and bunching windows are assumed to be symmetric for simplicity. We test how sensitive the estimated bunching elasticity is with regard to the window width c.

The control group is defined as those who have the majority of their trips in 2014 within a pre-specified "control window" $[c+d, c+d+w]$. OD pairs in the control window are at least d km longer than the distance cutoff but shorter than the next distance cutoff. The control group faces

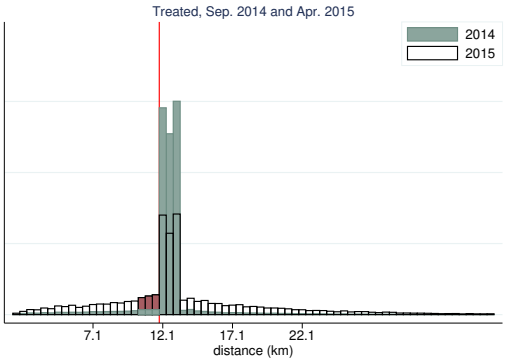
[2] In this analysis, we focus exclusively on frequent subway riders, defined as those who had three or more subway trips in the week of September 14, 2014. Assuming frequent riders have a stronger incentive to bunch, we postulate that the estimates are likely the upper-bound of e^b. The same methodology applies to all subway users, and the results are similar.

the same price as the treated group but is assumed to have no incentive to bunch because doing so would require them to make large trip adjustments. d is picked such that users in the control group have no plausible incentive to bunch, while $c+d+w$ is still short of the next distance cutoff. In the baseline, we choose $d = 2$ km. We define the control group's corresponding "bunching window" as $[c + d - w, c + d)$. With $w \leq d$, the bunching window for the control group lies entirely to the right of the cutoff and has the same price as a trip in the control window. Users in the control group thus have no incentive to shorten their trips to this bunching window.

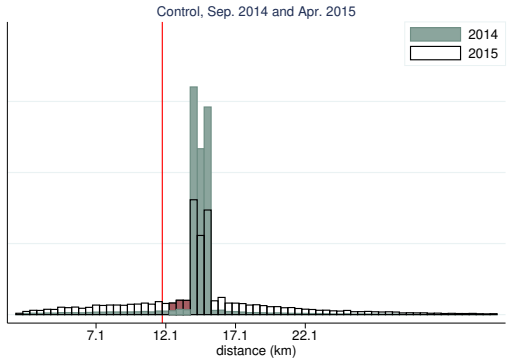
p_l is the fare of OD pairs to the left of the distance cutoff c and p_r is that to the right of the cutoff. $(p_r - p_l)/p_l$ thus captures the percent change in fare across the distance cutoff. N^r_{i14} is the number of trips from a treated (control) rider i in 2014 that falls in the *treated* (*control*) window, and ΔN^l_i is the change in the number of trips in the corresponding *bunching* window between 2014 and 2015.[3] $\Delta N^l_i / N^r_{i14}$ measures the percent change in the number of trips in the bunching window relative to the initial number of trips in the treated (control) window. The control group has no incentive to bunch, although this term may not necessarily be zero due to idiosyncratic shocks to the demand for various trips. The treated group is assumed to experience idiosyncratic shocks to demand that are drawn from the same distribution. In addition, users in the treated group have incentives to bunch, driven by the percent price difference across the distance cutoff, which is $(p_r - p_l)/p_l$. The coefficient associated with $T_i \cdot (p_r - p_l)/p_l$ can be interpreted as the bunching elasticity.

Figure B.3: Illustration of Estimating Bunching Elasticity

Panel A: Treated group



Panel B: Control group



Note: The graphs show the distribution of ridership by distance bin around the distance cutoff of 12 km. The distributions of the treated group and the control group are plotted separately. Blue bars indicate the ridership in 2014, and the white bars indicate the ridership in 2015.

Figure B.3 graphically illustrates the intuition of this strategy to estimate e^b. Focusing on the

[3]The number of trips in the week in September 2014 is converted to the number of a full month by multiplying 30/7.

distance cutoff at 12 km, the graphs plot ridership distribution by distance bins for the treated and the control groups and separately for 2014 and 2015. The bandwidth w is chosen at 1.5 km, and the distance between the treated window and the control window, d, is chosen at 2 km. The green bars in each figure show the number of trips in 2014, which by design are centered around the designated treated and control windows. Naturally, these users also have a small number of trips outside the designated windows.

The distribution of trips of the same users in 2015 is shown in white bars. Because many users, especially those who ride the subway frequently, have regular trip patterns, so the white bars are still concentrated around the designated window. However, due to idiosyncratic demand shocks, the distribution of the white bars is more dispersed than that of the green bars. We have selected the treated and control groups based on their travel patterns before the fare adjustment. The flattening out of the distribution is a natural regression toward the mean. It highlights the need for a control group to flesh out the bunching behavior in the treated group, under the key assumption that the distributions of the idiosyncratic shocks to the demand for subway trips are the same between the treated group and the control group. In other words, if the treated group had no incentive to bunch, we assume that the changes in the distribution of trips among the treated group are the same as those among the control group.

We are interested in testing whether the treated group has an abnormal cluster of trips that fall in the bunching window to the left of the cutoff, net of the natural dispersion due to idiosyncratic shocks. The change in the number of trips in the bunching window of the control group captures the natural dispersion driven by idiosyncratic shocks. Essentially, we compare the differences in the red areas in the two graphs shown above. This constitutes a difference-in-differences style estimation.

Table B.3: Bunching Elasticity

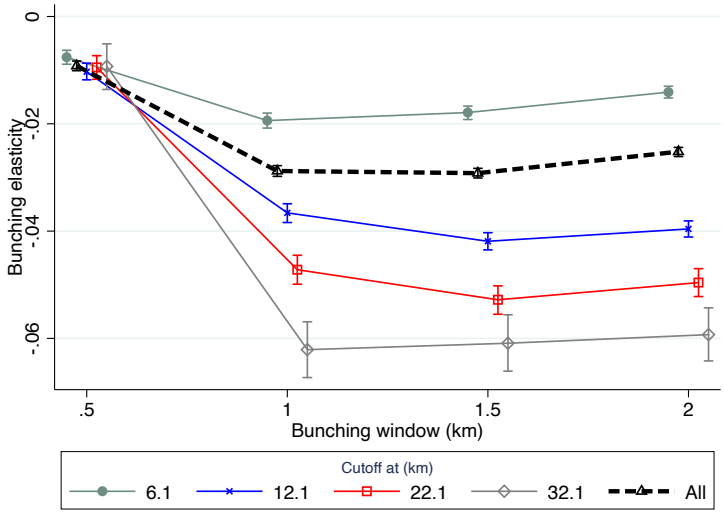
	(1)	(2)	(3)	(4)	(5)
	at distance cutoff (km)				
	6	12	22	32	all cutoffs
e^b	-0.018	-0.042	-0.053	-0.061	-0.029
	(0.001)	(0.001)	(0.001)	(0.003)	(0.000)
N (mil)	4.04	4.19	2.02	0.56	10.80

Note: Estimations of Equation B.1 are reported in Columns 1 through 4. The estimation of Equation B.2 is reported in Column 5. Each observation is a user. Weights are the number of each user's rides in the week of September 15, 2014, in the assigned window. The window width is 1.5 km. Standard errors are clustered at the user level.

The first four columns of Table B.3 report the estimation results with a bandwidth $w = 1.5$ km at the first four distance cutoffs. Bunching elasticity is generally small and somewhat larger at higher distance cutoffs. We experiment with different bandwidths w, and the results are largely

stable once the bandwidth is above 1 km, as is shown in Figure B.4. This is consistent with evidence from Figure 2, which suggests that most bunching is within 1 km from the distance cutoffs.

Figure B.4: Bunching Elasticity - Varying Window



Note: Estimations of Equation B.1 with varying bunching window w at each distance cutoff as well as jointly with all cutoffs. Vertical bars indicate 95% confidence intervals.

One single bunching elasticity combining all cutoffs can be estimated with the following equation:

$$\frac{\Delta N_i^l}{N_{i14}^r} = e^b \cdot \frac{p_{rc} - p_{lc}}{p_{lc}} \cdot T_{ic} + \theta_c + \varepsilon_{it}. \tag{B.2}$$

The outcome variable is the same as that in Equation B.1. T_{ic} indicates treatment status of card i at cutoff c. $(p_{rc} - p_{lc})/p_{lc}$ is the percent change in subway fare from the left to the right of cutoff c. We include cutoff fixed effects θ_c to account for the heterogeneity in ridership changes at different cutoffs. The estimated bunching elasticity is -0.029 with a bandwidth of 1.5 km, which is similar in magnitude to the weighted average of bunching elasticity estimated separately for each cutoff. Again, the result is not sensitive to w as long as it is above 1 km. Take $e^{RD} = -0.387$ (Table 2 Column 5) and $e^b = -0.029$, we recover the demand elasticity as $e = e^{RD} - 2e^b = -0.33$. Bunching makes the elasticity larger (in magnitude) by 18%. This estimated demand elasticity is quantitatively similar to the OD pair RD approach result (-0.36 from Table 2 Column 6), which adjusted for strategic bunching by cutting out a donut hole.

B.3 Demand Elasticity from User-level Estimations

The ridership data across different waves can be linked via an anonymized card ID. Tracking the same user before and after the fare structure change can be another way to identify the demand elasticity. In this subsection, we show that this approach, though intuitive and appealing at first glance, suffers from some fundamental identification challenges.

We estimate the following equation using ridership data in September 2014 and April 2015.

$$\Delta \ln D_i = e^{\text{card}} \cdot \Delta \ln exp_i + f(\ln Dist_{i,t_0}) + g(\ln Rides_{i,t_0}) + \Delta \varepsilon_i \tag{B.3}$$

Each observation is a unique IC card (a user) that showed up in the September 2014 data, which covers the entire week between the 15^{th} and the 20^{th}. D_i is the change in ridership between the two periods. Two variables are used to proxy for demand. The first is the total number of trips the user took during the sample period. The second is the total distance of those trips. We convert monthly ridership in April 2015 into weekly by multiplying 7/30. We add one to both the number of trips and the total distance to avoid the problem of taking logarithm over zero. The results are similar if inverse hyperbolic sine is used instead of the logarithm, or using levels as the outcome variable and then converting the coefficient into elasticity.

$\Delta \ln exp_i$ is the log change in expenditure for user i if she is to take the same bundle of trips she took in September 2014 after the fare adjustment. We measure the change in expenditure based on the weekly data in both waves and do not take into consideration potential discounts users qualify for. It is a necessary assumption because a user's full-month ridership cannot be simply extrapolated from one week of data we have for the pre-adjustment period. It is arguably not an unrealistic assumption because, as shown in the paper, consumers mostly react to the listing price and do not respond to discounts in any way.

e^{card} can be interpreted as the demand elasticity. Because we are interested in the average demand elasticity of all trips, we weigh each user by her initial ridership in 2014. When the outcome variable is the log change in the number of trips, the weight is the number of trips in the initial period; when the outcome variable is the log change in total distance, the weight is the total distance in the initial period.

$f(.)$ and $g(.)$ are flexible polynomials of log distance and log number of trips in September 2014. Conditional on the number of trips a user takes, she will experience a larger increase in her expenditure if she rides longer distances. Controlling flexibly for the total distance alongside the number of trips, $\Delta \ln exp_i$ captures the non-linearity in pricing as a function of distance. We pick $f(.)$ and $g(.)$ to have an up-to-5^{th} order polynomial, and estimations are generally robust to the order of polynomials.

Compared with the OD-pair RD approach in the baseline, this approach is immune to strategic

bunching in trip distance. If a user chooses to shorten her trip to bunch below the distance cutoff, the trip is still recorded in her tally, and the ridership measured as total distance is only mildly smaller. In addition, this approach can estimate the demand elasticity measured in distance, which some may argue is a better measure of subway usage than the number of trips.

Equation B.3 maintains an implicit assumption that the ridership patterns observed in September 2014 measure the *true* demand for subway trips. There are at least two reasons why this may not be the case. First, one week of ridership data can be a less-than-perfect approximation of one's true needs for subway rides. Second, demand for subway trips may change over time for reasons other than the fare structure change. For example, in the full month data in April 2015, during which the fare structure was unchanged, there was substantial week-to-week variation in ridership within the same user.

Both reasons essentially reflect the classical measurement error problem. Ordinary least squares (OLS) regressions with classical measurement error in the explanatory variable suffer from the attenuation bias. The conventional approach to address the problem is to find another proxy for the underlying variable. The other proxy may also have measurement errors, but as long as the measurement errors from the two proxies are not correlated, one can be used as an instrumental variable for the other.

One idea along this line is to further split the one week of data in September 2014 into two halves and generate two measures of the underlying demand. For example, one measure can be constructed using the ridership pattern on Monday, Wednesday, Friday, and Sunday, extrapolated to the whole month, while the other measure is constructed using the remaining days in the week. Yet, this is apparently asking too much from a single week of data; the first stage is weak.

Thus, we estimate Equation B.3 using OLS. In light of the discussion above, we postulate that the OLS models likely underestimate the true demand elasticity.

Column 1 of Table B.4 reports the demand elasticity for the number of trips is -0.27. It is slightly smaller than the baseline result from the OD-pair RD regressions, which is consistent with the attenuation bias due to classical measurement error. This number can be seen as the lower bound of the demand elasticity unaffected by the incentive to bunch below the distance cutoffs. Column 2 reports the demand elasticity in terms of passenger-miles is -0.51. The elasticity is larger in magnitude than that for the number of trips. It is also expected, because longer trips experienced larger fare rises and were reduced by a larger share. If the elasticity for the number of trips is the same across rides of different distances, the impact on passenger-mile is larger than that on the number of trips.

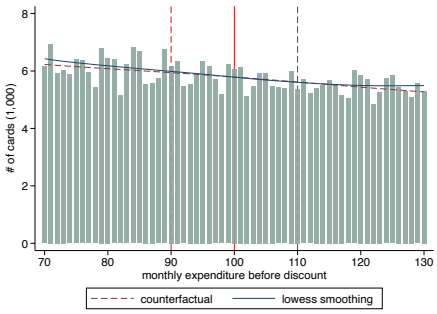
Table B.4: Demand Elasticity from Card Level Estimations

	(1)	(2)
	$\Delta \ln D_i$ in	
	# of trips	total distance
$\Delta \ln p_i$	-0.269	-0.508
	(0.010)	(0.017)
$f(.), g(.)$	5^{th}-order polynomial	

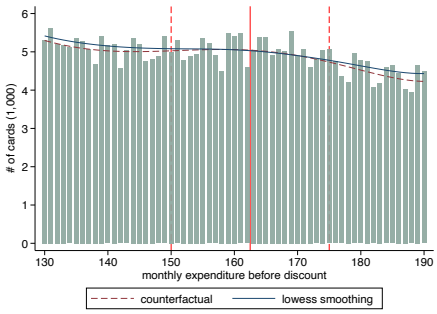
Note: The sample includes the 6.7 million users that appeared in the week of September 15, 2014. Each observation is a user. Estimations are based on Equation B.3. The dependent variable is the log change in the number of trips (Column 1) or total distance (Column 2) between the week of September 15, 2014, and April 2015. Both dependent variables are replaced with plus one to avoid taking logs over zeros. Up to 5^{th} order polynomials of the log ridership and the log distance in the week of September 15, 2014, are controlled for in all regressions. In Column 1, each observation is weighted by the number of trips taken by each user in the initial period. In Column 2, each observation is weighted by the total distance from all trips during the initial period. Robust standard errors are reported in parentheses.

Figure B.5: Density Distribution around Budget Kinks: Regular Commuters

Panel A: First Kink at 100 *yuan*



Panel B: Second Kink at 162.5 *yuan*



Note: Data are from the full-month ridership records in April 2015. The sample includes frequent users who have a regular travel pattern. The discount schedule creates three kinks in the budget line, here we show the distribution of pre-discount expenditure at the first two kinks at 100 and 162.5 yuan, respectively. In both graphs, the solid vertical line indicates the kink point, the two dashed vertical lines indicate the neighborhood that is excluded when we impute the counterfactual density. The dashed red line depicts the counterfactual distribution fitted by a polynomial excluding the neighborhood around the kink point. The blue line depicts the smooth fitted line with the neighborhood included. Fitted density and actual density is imputed from estimating Equation 2. The non-convex kinked budget would imply the actual distribution (green bars) to be below the counterfactual distribution in the narrow neighborhood around the kink point, which is evidently not present in the graph.

B.4 Test for Rationality with Regular Commuters

In Section 3.3.2, we show there is no evidence that consumers rationally or quasi-rationally respond to the non-linear monthly budget. Notice that users represented in Figure 6 are frequent subway riders with a monthly expenditure near or above the expenditure cutoff to qualify for discounts. For this group of users, the discounts could account for a non-negligible share of their monthly budget. One may suspect that frequent users with a regular travel pattern are more

likely to be able to predict their monthly demand and rationally respond to it. Figure B.5 repeats the same density analysis among those we classify as regular commuters. Again, the empirical monthly budget has no visual or statistical evidence of hollows or dents around the first two non-convex kinks. We do not show a similar graph around the third kink point (at the pre-discount monthly expenditure of 662.5 yuan) because few users are in the neighborhood.

B.5 Additional Tests for Myopia

Section 3.3.3 shows that evidence for myopia is limited. Here, we present some additional evidence.

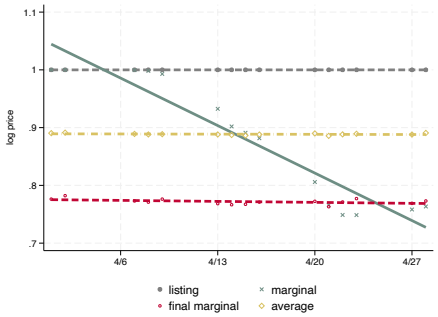
Most users do not spend enough to qualify for any discount. Distinguishing different behavioral assumptions is irrelevant for them. First, we take trips from less-frequent users (those with a pre-discounted monthly expenditure of less than 70 yuan) and estimate Equation 4. This is a placebo test that checks whether there is any secular trend in demand elasticity during the course of the month that could confound with the temporal pattern. The series in gray squares in Figure B.6 Panel C represents $\hat{e}_t^{\text{listing}}$ from those estimations. These estimates follow a near-perfect flat line. The fitted linear line has a slope of -0.00018 and a robust standard error of 0.001. This is evidence that the underlying demand elasticity, as a structural parameter, does not change over time.

We then focus on trips from frequent users with a pre-discounted monthly expenditure of more than 70 yuan. We include some users who spend less than 100 yuan (the lowest cutoff to qualify for any discount) for two reasons. First, even if they do not spend enough to qualify for discounts, these users spend enough to get close to the cutoff, and they may expect and respond to the discount somehow. Second, the inclusion of those users introduces variation and helps with identification. To see that, imagine we only keep users whose monthly expenditures are above the threshold and are within a narrow range. These users have similar expenditures and qualify for similar discounts; they may take trips that land at different sides of the distance cutoffs. So, there will be no regression discontinuity in discounts around the distance cutoff. We also consider a subgroup of frequent users with a regular commuting pattern (as identified from the $K-$means clustering of travel patterns). One may argue that users with a regular travel pattern are more likely to be aware of the discounts and have a stronger incentive to respond to them.

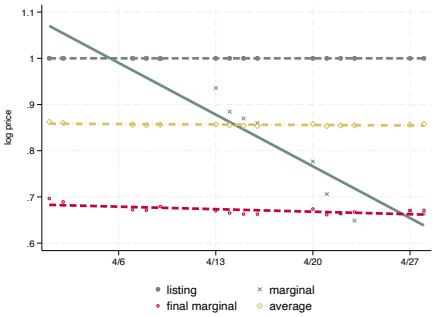
The first two panels of Figure B.6 plot the estimates from regression discontinuities in prices, using Equation 3. Panel A corresponds to frequent users, and Panel B corresponds to regular commuters. While discontinuities in the listing price, month-end marginal price, and monthly average price all lie in flat lines for both groups, discontinuities in the instantaneous marginal price are in downward trend lines. Panel C reports $\hat{e}_t^{\text{listing}}$ from estimating Equation 4 for both

Figure B.6: Tests for Myopia by Sub-samples

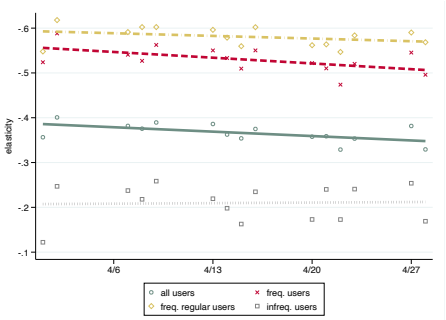
Panel A: RD in Prices, Frequent Users



Panel B: RD in Prices, Frequent Regular Users



Panel C: Elasticity Implied by Listing Price



Note: Panels A and B plot regression discontinuity estimates of perceived prices under various behavioral assumptions by each day between Monday and Thursday in the month of April 2015. The regression equation is described in Equation 3. In Panel A, the sample includes trips from frequent users with pre-discounted monthly subway expenditures of more than 70 yuan. The sample in Panel B includes trips from frequent users with regular commuting patterns. In Panel A, the linear fitted line for $\hat{\rho}_t^{\text{myopic}}$ has a slope of -0.0117 and a robust standard error of 0.0012. In Panel B, the linear fitted line for $\hat{\rho}_t^{\text{myopic}}$ has a slope of -0.0160 and a robust standard error of 0.0017. Panel C plots regression discontinuity estimates of demand elasticity by date in the month of April 2015, assuming consumers respond only to the listing price. The regression equation is described in Equation 4. Each of the four series corresponds to a different sample: (1) trips from all users, (2) trips from frequent users, (3) trips from frequent users who have a regular commuting pattern, and (4) trips from less frequent users. Lines in the corresponding color are linear fits of those coefficients. The fitted line has a slope of 0.0014 and a robust standard error of 0.0007 for trips from all users; it has a slope of 0.0018 and a robust standard error of 0.0009 for trips from frequent users; it has a slope of 0.0008 and a robust standard error of 0.0009 for trips from frequent and regular commuters; it has a slope of -0.00018 and a robust standard error of 0.0017 for trips from infrequent users. Estimates are weighted by the number of trips in the OD pair in September 2014. Standard errors are clustered at the OD-pair level. Confidence intervals are suppressed in both panels for clean illustration.

groups. Red dots represent estimates from frequent users, while yellow diamonds represent those from frequent and regular users. Both series of $\hat{e}_t^{\text{listing}}$ show a slight downward trend over the course of the month, consistent with some evidence of users being myopic. But the magnitude of the downward sloping is small. The fitted line has a slope of 0.0018 (robust standard error of 0.009) for the frequent users as a whole and a slope of 0.0008 (robust standard error of 0.0009)

for the subset of regular commuters. If users are all myopic, declines in the discontinuity in the instantaneous marginal price shown in Panels A and B would imply a 33.7% decline in $\hat{e}_t^{\text{listing}}$ over the course of the month for frequent users and a 44.8% decline for regular commuters. In fact, the decline in $\hat{e}_t^{\text{listing}}$, as shown in Panel C, is 10% for the former and 4.3% for the latter. The results lead to the conclusion that users in our sample cannot be described as myopic to the first order.

B.6 Additional Results on the Composition of Behavioral Types in Response to Non-linear Budget Constraints

Appendix Table B.5 reports the results from estimating versions of Equation 5 and Equation 7 using OLS. The estimates are remarkably similar to those in Table 4, which are estimated using the Two-stage Least Squares estimator where discounted prices are instrumented using counterparts constructed to purge out instantaneous shocks to the demand for subway trips.

Table B.5: Mixture Model and the Composition of Behavioral Types (OLS)

	dep var: $\Delta\ln(N_{od,t})$							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Log listing p	-0.289	-0.882	-0.578	-0.630	-0.599	-0.736	-0.655	-0.714
	(0.192)	(0.191)	(0.081)	(0.183)	(0.044)	(0.073)	(0.055)	(0.079)
Log instan. marginal p	0.049	-0.001	0.026	-0.003	0.031	-0.002	0.072	0.029
	(0.042)	(0.028)	(0.032)	(0.028)	(0.013)	(0.012)	(0.009)	(0.008)
Log final marginal p	-0.064	-0.176	-0.017		-		-	
	(0.126)	(0.069)	(0.063)		-		-	
Log avg. p	-0.048	0.506		0.071		-		-
	(0.284)	(0.226)		(0.185)		-		-
Polynomials of								
log final marginal p					X		X	
log avg. p						X		X
e constrained at					-0.569	-0.561	-0.531	-0.525
Sample	all	frequent					freq. and regular	
N (mil.)	1.47	1.39					1.20	

Note: The table reports results from estimating various versions of the mixture model using the OLS. Each observation is an OD pair. The dependent variable is the log difference between the ridership from the specific sample of users in that OD pair on a day of April 2015 and the ridership from the corresponding user group and on the same day of the week in September 2014. For Column 1, the sample includes ridership from all users. In Columns 2 through 6, the sample includes trips from frequent users with a monthly expenditure of more than 70 yuan. In Columns 7 and 8, the sample includes trips from frequent users who have a regular commuting pattern. Columns 5 through 8 account for optimization errors by including either a 5^{th}-order polynomial of log final marginal price (Columns 5 and 7) or that of log average price (Columns 6 and 8). In those regressions, the overall elasticity is constrained to be that estimated in the corresponding specification in which optimization errors are not accounted for. All regressions include a 5^{th}-order polynomial of OD-pair distance. Standard errors are two-way clustered at the origin and destination stations.

Table B.6: Pairwise Mixture Models

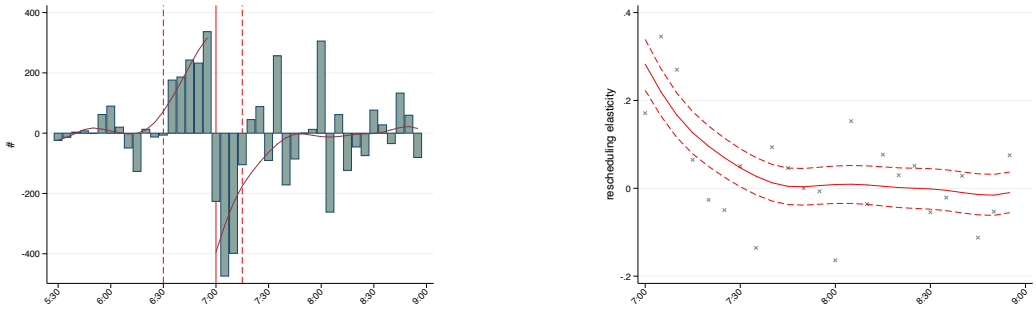
	dep var: $\Delta\ln(N_{od,t})$								
Panel A: OLS	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Log listing p	-0.313	-0.292	-0.193	-0.582	-0.563	-0.629	-0.558	-0.529	-0.574
	(0.090)	(0.078)	(0.198)	(0.070)	(0.073)	(0.184)	(0.072)	(0.077)	(0.174)
Log instan. marginal p	-0.035			0.015			0.031		
	(0.089)			(0.060)			(0.040)		
Log final marginal p		-0.062			-0.007			-0.004	
		(0.079)			(0.069)			(0.055)	
Log average p			-0.164			0.068			0.049
			(0.209)			(0.196)			(0.179)
Panel B: IV	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Log listing p	-0.316	-0.289	-0.192	-0.586	-0.565	-0.636	-0.556	-0.523	-0.561
	(0.090)	(0.078)	(0.197)	(0.071)	(0.073)	(0.183)	(0.072)	(0.077)	(0.174)
Log instan. marginal p	-0.032			0.019			0.028		
	(0.089)			(0.060)			(0.040)		
Log final marginal p		-0.064			-0.005			-0.014	
		(0.078)			(0.069)			(0.056)	
Log average p			-0.165			0.075			0.033
			(0.208)			(0.194)			(0.179)
Sample	all			frequent			freq. and regular		
N (mil.)	1.47			1.39			1.20		

Note: The table reports results from estimating pairwise mixture models where the assumption of consumers being oblivious to quantity discounts is cast against an alternative behavioral assumption. Each observation is an OD pair. The dependent variable is the log difference between the ridership from the specific sample of users in that OD pair on a day of April 2015 and the ridership from the corresponding user group and the day of the week of September 2014. In Columns 1 through 3, the sample includes ridership from all users. In Columns 4 through 6, the sample includes trips from frequent users with a monthly expenditure of more than 70 yuan. In Columns 7 through 9, the sample includes trips from frequent users with regular commuting patterns. All regressions include a 5^{th}-order polynomial of OD-pair distances. Panel A estimates the model using the OLS. In Panel B, log instantaneous marginal price, log final marginal price, and log average price are instrumented using counterparts that replace the same-day actual ridership with predicted ridership. Standard errors are two-way clustered at the origin and destination stations.

Appendix Table B.6 reports estimation results from pairwise mixture models in which the listing price is cast against each alternative price. The model specification follows Equation 5. Three samples are considered: (1) the full sample including trips from all users, (2) trips from frequent users, and (3) trips from frequent users who also have a regular travel pattern. There is overwhelming evidence for all three user samples that users respond only to the listing price. In all estimations, the implied demand elasticity remains consistent within the corresponding sample of users. Both OLS and IV estimations yield quantitatively similar results.

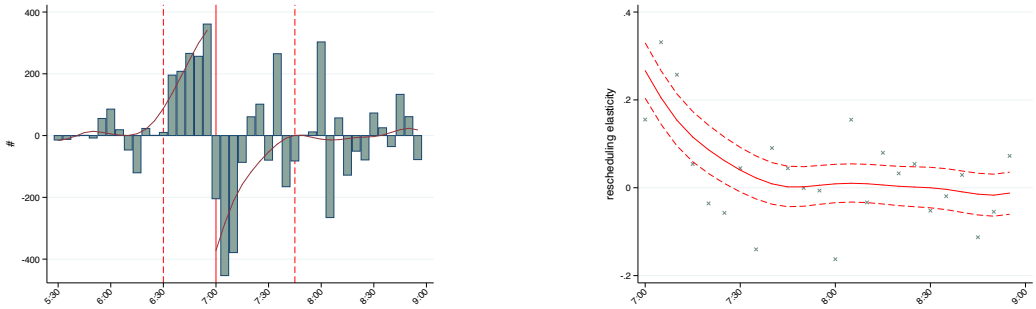
Figure B.7: Robustness Estimations of the Rescheduling Elasticity - Varying Bunching and Missing Windows

Panel A: Rescheduling Window between 6:30 AM and 7:15 AM



B=1,168, M=1,154, Counterfactual number of trips in the missing window=13,148

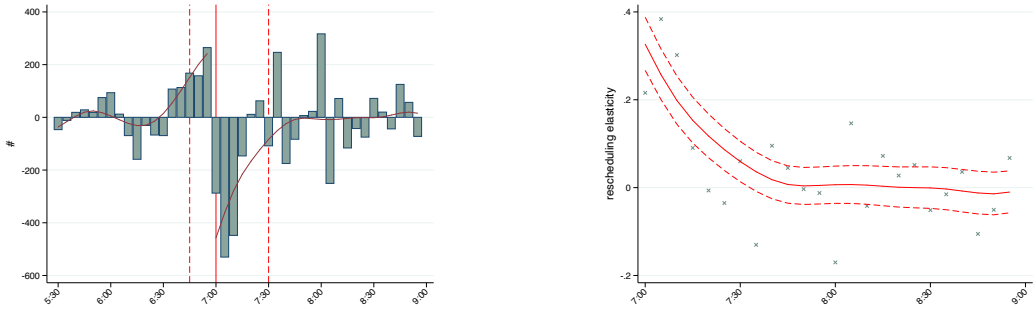
Panel B: Rescheduling Window between 6:30 AM and 7:45 AM



B=1,298, M=1,028, Counterfactual number of trips in the missing window=42,924

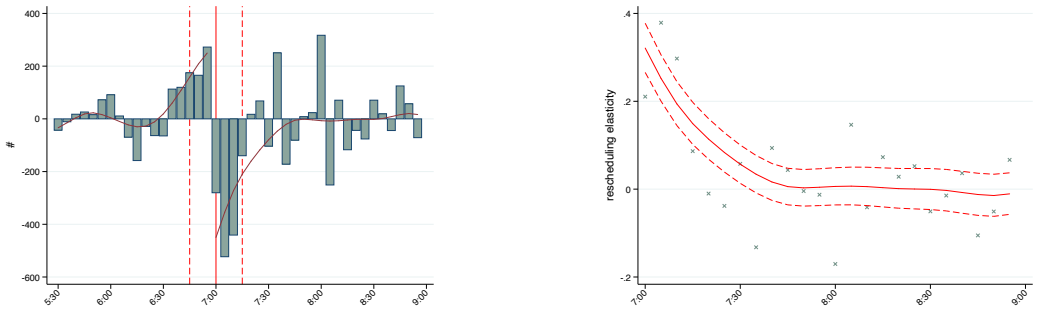
Panel C: Rescheduling Window between 6:45 AM and 7:30 AM

Missing Mass and Bunching | Rescheduling Elasticity



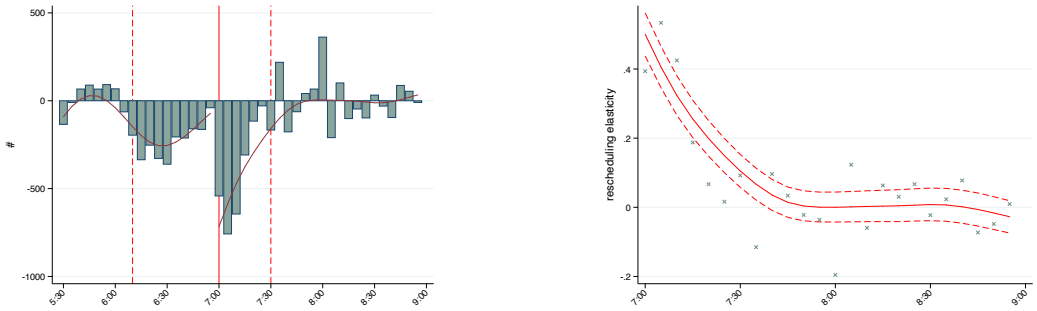
B=590, M=1,372, Counterfactual number of trips in the missing window=27,807

Panel D: Rescheduling Window between 6:45 AM and 7:15 AM



B=613, *M*=1,307, Counterfactual number of trips in the missing window=13,301

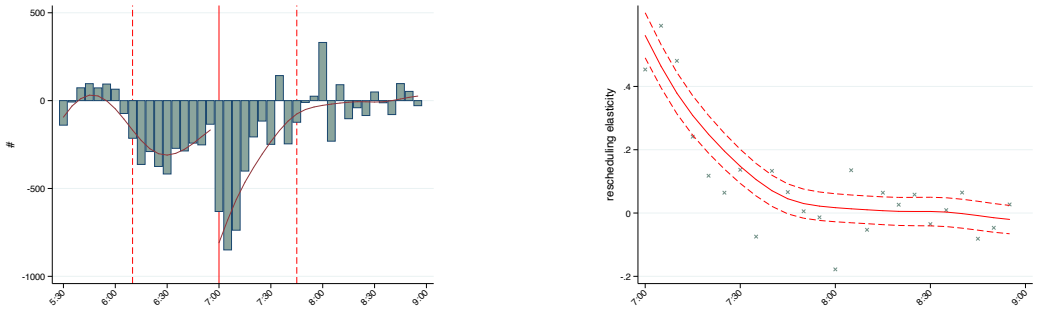
Panel E: Rescheduling Window between 6:15 AM and 7:30 AM



B=-2,059, *M*=2,446, Counterfactual number of trips in the missing window=28,882

Panel F: Rescheduling Window between 6:15 AM and 7:45 AM

Missing Mass and Bunching | Rescheduling Elasticity



B=-2,632, *M*=3,393, Counterfactual number of trips in the missing window=45,289

Note: Estimation of the rescheduling elasticity with the demand elasticity imposed as -0.36. See notes to Figure 8 for details.

B.7 Robustness Estimations of the Rescheduling Elasticity

Varying the width of the rescheduling window. This section provides robustness checks to estimate the rescheduling elasticity introduced in Section 3.4. The first set of robustness checks

involves varying the width of the rescheduling window. While a wide rescheduling window can capture rescheduled trips far away from the cutoff, it comes at the cost of a smaller sample to fit the counterfactual ridership curve, resulting in a less precise estimation.

Figure 8 Panel C shows that setting the rescheduling window to be 30 minutes around the time cutoff is sufficient to capture rescheduled trips. Here, we test how sensitive the results are to the choice of window width, in particular, (1) whether the missing mass M remains approximately equal to the bunching mass B, and (2) whether the estimated rescheduling elasticity remains stable.

We first hold the bunching window unchanged but alter the missing window to be between 7 and 7:15 AM (Figure B.7 Panel A) and between 7 and 7:45 AM (Panel B). The missing and bunching masses are largely unchanged from the baseline. This is unsurprising because we have shown that the EBD affects few trips beyond 7:15 AM. The corresponding rescheduling elasticity curves are also essentially unchanged.

The next two panels show results where the bunching window is narrowed to between 6:45 and 6:59 AM. Evidence shows that some bunching trips are placed before 6:45 AM, and the bunching mass is smaller than the missing mass. However, the counterfactual ridership curve is still robustly estimated, and the corresponding elasticity curves, which are defined for the time window after 7 AM, are little changed.

We expand the bunching window to between 6:15 and 6:59 AM in Panels E and F. While a wider bunching window allows for ample space for rescheduled trips to land, it leaves relatively few observations to the left of the bunching window for the estimation of the counterfactual curve. The counterfactual curve fits the data poorly and generates a negative value for the bunching mass.

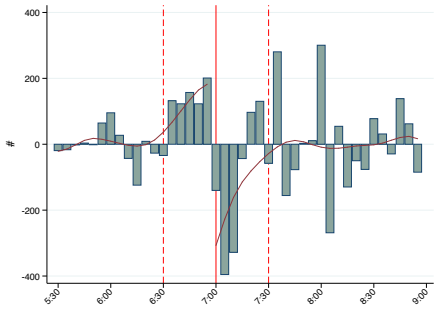
Joint estimation of demand and rescheduling elasticities. The baseline and robustness checks presented in Figure B.7 all impose a demand elasticity that is estimated from the OD pair regression discontinuity design. However, the demand elasticity for the relevant population in the EBD experiment may differ. Here, we present an approach that jointly estimates both the demand and rescheduling elasticities.

Instead of imposing demand elasticity, we start with an initial guess and conduct the first three steps as described in Section 3.4. By the end of Step 3, we check whether the bunching mass B and the missing mass M are sufficiently close. If not, the demand elasticity is updated to $e^{d,1} = (B/M) \cdot e^{d,0}$, where $e^{d,0}$ is the initial guess of demand elasticity, and return to Step 1. We repeat this procedure until B and M converge (we set the criterion as being equal to each other by the integer), and the rescheduling elasticity is calculated according to Equation 8.

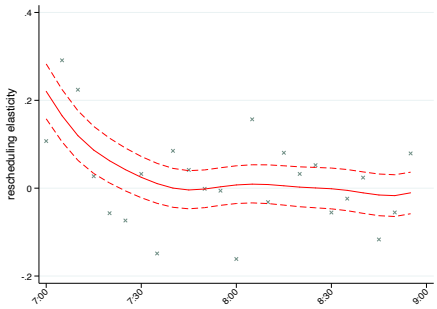
We set $e^{d,0}$ to be -0.36, and the iterations converge when e^d is -0.37, which is very close to the baseline demand elasticity. Missing and bunching masses converge at 1,172. Figure B.8 shows

Figure B.8: Joint Estimation of Demand and Rescheduling Elasticities

Panel A: Missing and Bunching Masses



Panel B: Rescheduling Elasticity

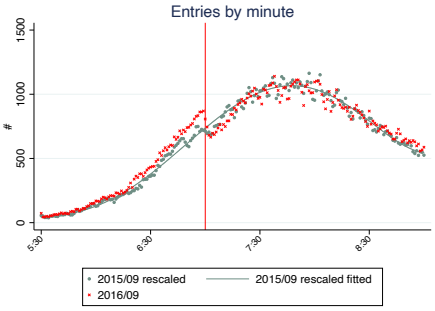


Note: The data include all trips in the week of September 12, 2016 (Monday through Friday). The sample includes trips that start between 5:30 AM and 9 AM and originated from 16 stations that had the EBD implemented in December 2015. Panel A shows the excessive bunching (before 7 AM) and the missing mas (after 7 AM) in five-minute bins. Demand and rescheduling elasticities are jointly estimated, with the total bunching and missing masses restricted to be equal. The red curves are non-parametric fits for the size of the missing mass and that of the excessive bunching, respectively. Panel B reports the associated rescheduling elasticity by five-minute bins. Smoothed 95% confidence intervals (the smoothed 2.5th percentile and the smoothed 97.5th percentile from 1,000 bootstraps) are in dashed lines.

that the distributions of the bunching and missing masses (Panel A), as well as the rescheduling elasticities (Panel B), are also very similar to the baseline.

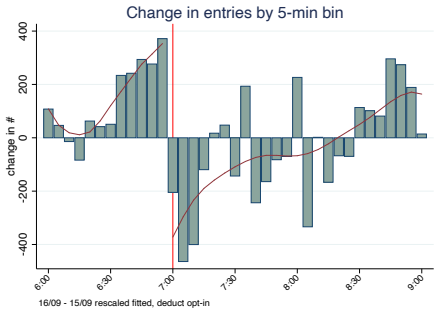
Figure B.9: Robustness Checks for Rescheduling Elasticity - Using September 2015 as Control

Panel A: Entry by Minutes: Sep 2015 vs. Sep 2016

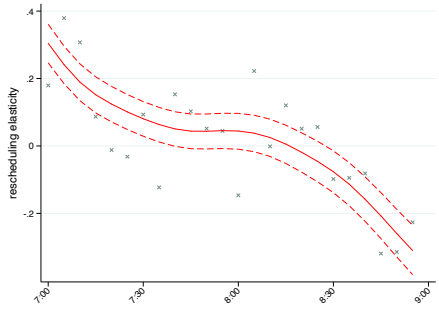


Panel B: Impose Demand Elasticity

Missing Mass and Bunching



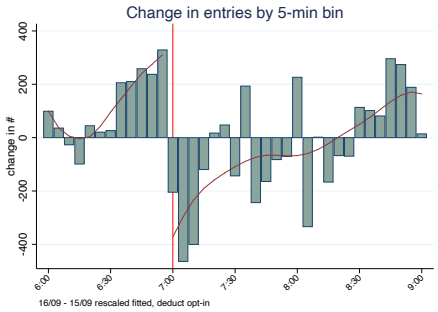
Rescheduling Elasticity



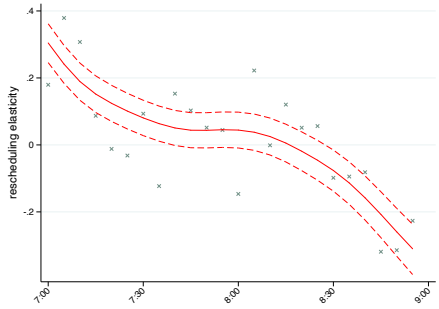
B=1468, M=1267, Counterfactual number of trips in the missing window=31,710.

Panel C: Joint Estimation of Demand and Rescheduling Elasticities

Missing Mass and Bunching



Rescheduling Elasticity



B=1,267, M=1,267, Counterfactual number of trips in the missing window=31,710.

Note: The figure shows the alternative estimation of the rescheduling elasticity where the ridership in September 2015 is used as the counterfactual. See notes in Figure 8 for details.

Using ridership in September 2015 as control. So far, we have been using data from the week of September 12, 2016, a time when the EBD was already implemented. We impute the counterfactual ridership without the EBD by fitting a smooth function of time while purging out opt-in trips and carving out a time window for rescheduled trips. As a natural control group, we also have ridership data from the week of September 15, 2015, at the same time of the year but before the EBD's implementation.

We calculate the total number of entries by minute in the same 16 stations on the weekdays of the week of September 15, 2015. The aggregate ridership may have changed between the two years, so we rescale the number of trips in September 2015 between 7:30 and 9 AM to match that in September 2016. The time window for the rescaling is chosen to exclude the impacts of opt-in trips and rescheduled trips, with the opt-in window between 5:30 and 6:59 AM and the rescheduling window assumed to be between 6:30 and 7:29 AM. The assumption is that the secular percent changes in ridership without the EBD are the same for each minute between the sample time window (5:30-9 AM).[4] The numbers of entries by minute in September 2016 are plotted in red dots in Panel A of Figure B.9, while the rescaled numbers of entries in September 2015 are plotted in green crosses. There is a visually clear discontinuity around 7 AM for the former but no such discontinuity for the latter.

We fit the rescaled number of entries in September 2015 during the sample time window using a flexible smooth function of time t. The fitted curve, shown in green solid line in Panel A, serves as the counterfactual number of entries in September 2016.

We then add opt-in trips to the counterfactual ridership between 5:30 and 6:59 AM. This is done in two different approaches. The first approach imposes the baseline demand elasticity, -0.36. The second approach aims at estimating the demand and rescheduling elasticities jointly. We start with an initial guess of the demand elasticity (for which we use the baseline estimate) and iterate it until the resulting bunching mass is sufficiently close to the missing mass. The differences between the observed entries in September 2016 and the counterfactual entries adjusted for opt-in trips are missing and bunching masses. The rescheduling elasticity and confidence intervals are obtained following the baseline procedure described in the main paper.

Panel B of Figure B.9 plots the estimation results when we impose the demand elasticity. The panel on the left shows the missing and bunching trips in five-minute bins. Although we do not restrict the bunching mass to equal the missing mass, they come close to each other with $B = 1,468$ and $M = 1,267$. The panel on the right plots the associated rescheduling elasticity. The rescheduling elasticity is around 0.3 for trips just to the cutoff's right and declines quickly to zero by around 7:20 AM. Those estimates again correspond to a small share of peak-hour ridership

[4] In other words, we allow for proportional "shifts" in ridership by minute but do not allow for "rotations" in the shape of ridership as a function of time.

that is rescheduled due to the EBD.

Panel C reports the results from the joint estimation of demand and rescheduling elasticities. The estimated demand elasticity in this specification is around -0.4, which is again close to the baseline. Consequently, bunching and missing masses (left panel) and the associated rescheduling elasticity (right panel) are essentially unchanged from those in Panel B and those in the baseline.

In both approaches, however, the counterfactual ridership fits less well towards the end of the sample time window (after 8:30 AM). It suggests that percentage changes in ridership between September 2015 and September 2016 are not uniform across the sample period. Therefore, although using ridership in September 2015 as a control is a natural choice, and the results are similar, we prefer the baseline approach that uses only data from September 2016.

C Details of the Clustering Algorithm and User Classification

C.1 Additional Details of the K-means Clustering Algorithm

Section 3.2.1 describes a K-means clustering algorithm that classifies users by the overall usage and temporal and geographical patterns of their subway rides. Here, we provide additional details on the construction of the predictors and the algorithm.

We construct variables to describe the geographical pattern of users' trips. We say a user has trips that exhibit regular geographic patterns if a large fraction of her trips are between a small number of locations. We first divide the urban areas of Beijing (within the 6^{th} ring road) into 3,000 equal-sized location bins. Each bin covers an area of a little less than one square kilometer. Subway trips are then mapped into location bin pairs. Note that location bin pairs are not directional. A commuter who travels from bin A to bin B in the morning and then from bin B to bin A in the afternoon is regarded to have trips in the same location bin pairs.

Two predictors are constructed to describe the concentration of users' trips. First, we construct the Herfindahl–Hirschman index (HHI) of the location bin pairs. Second, we calculate the OD location bin concentration rate, which is the total number of trips a user takes during the month divided by the number of unique location bin pairs from those trips. Both the HHI and the concentration rate measure the geographic regularity of the user's trips. The HHI is widely used as an indicator for concentration, but it can be sensitive to the total number of trips. For example, if a user has only one trip in the month, her HHI will be one, indicating her geographic travel pattern is very regular. But one may argue that her pattern is less regular than a user who takes ten trips, nine of which are in the same location bin pair, although the latter will have an HHI of less than 1. The geographic concentration rate will say the former user has a value of one,

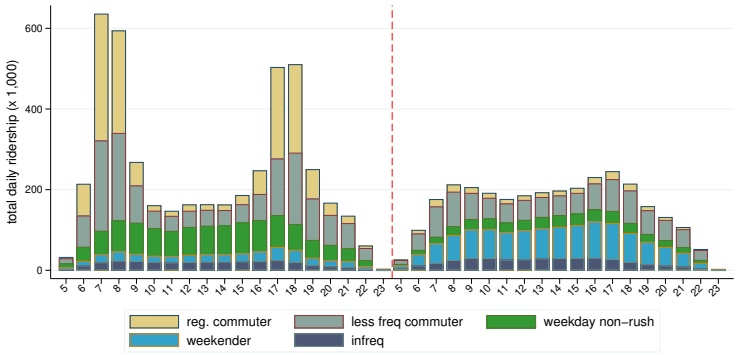
while the latter has a value of five. Combined with the total number of trips, which is also in the set of predictors, the algorithm is able to separate infrequent subway riders from frequent users who have regular travel patterns.

All predictors are first standardized, so they all have the same variation they bring to the algorithm. The algorithm also requires first specifying the number of clusters. We start with two (in addition to the infrequent users, separated out before running the algorithm). We gradually increase the number of clusters and inspect the characteristics of each group. The number of clusters is determined when allowing for a larger number of clusters does not lead to new unique travel patterns.

Trip characteristics by user types are reported in Table 3. User types are mostly identified by the number and timing of trips. Among frequent commuters, regular and less-regular commuters are separated by whether their trips are concentrated in a handful number of location bin pairs. We name each group with the most salient feature of their travel patterns.

C.2 Ridership by Time and User Type

Figure C.1: Trips Composition by Time and User Type



Note: Cards are classified into five categories based on their travel patterns in April 2015 using a K-means clustering algorithm. The graph shows the composition of trips by card type and by day and time.

Figure C.1 shows the ridership by hour and user type in April 2015. The aggregate ridership exhibits a salient twin-peaked pattern on weekdays. Ridership is highest during morning and afternoon peak hours (between 7 and 9 in the morning and between 5 and 7 in the afternoon). In each of the four peak hours, more than half a million passengers entered the subway system. Ridership during the weekend is more smoothly distributed throughout the day.

Trips belonging to each of the five user groups according to the K-means clustering algorithm are color-coded. Regular peak-hour commuters and less-regular commuters contribute to most of the ridership during peak hours. Weekday non-rush hour travels contribute a small proportion to peak-hour ridership but account for about half of the ridership during non-peak hours.

Weekenders take about one-third of all trips during the weekend. Trips from infrequent riders account for a small share of the aggregate ridership and are evenly spread out.

C.3 Classification of Users in September 2014

Table C.1: User Classification and Characteristics (Week of September 15, 2014)

	infreq. users	rush-hour commuters	less-reg. commuters	weekday non-rush	weekenders
	(1)	(2)	(3)	(4)	(5)
# of users (mil.)	3.17	0.77	1.02	1.01	0.67
# of rides (weekly)	1.53	8.31	9.84	4.96	4.28
	(0.50)	(3.12)	(5.11)	(2.15)	(1.54)
total distance (km)	24.94	131.46	152.37	76.07	66.06
	(18.66)	(91.24)	(105.20)	(50.99)	(40.21)
share of rides during					
weekday AM rush	0.14	0.43	0.36	0.14	0.07
	(0.29)	(0.22)	(0.18)	(0.17)	(0.13)
weekday PM rush	0.14	0.33	0.28	0.15	0.09
	(0.29)	(0.22)	(0.18)	(0.17)	(0.14)
weekday non-rush	0.36	0.15	0.21	0.61	0.16
	(0.44)	(0.19)	(0.17)	(0.21)	(0.17)
weekend	0.36	0.09	0.15	0.10	0.68
	(0.50)	(0.13)	(0.14)	(0.14)	(0.22)
# of weekdays traveled	0.76	4.27	4.31	2.44	0.96
	(0.58)	(0.98)	(0.87)	(0.94)	(0.74)
location bin HHI	0.91	0.96	0.48	0.47	0.48
	(0.19)	(0.09)	(0.17)	(0.24)	(0.32)
OD location bin concen. rate	1.35	7.21	2.90	1.93	1.78
	(0.48)	(2.83)	(1.17)	(1.07)	(0.94)

Note: Cards are classified into five categories based on their travel patterns in the week of September 15, 2014 using a $K-$means clustering algorithm. The table reports the summary statistics of travel patterns for each card category.

We apply the K-means clustering algorithm to users that showed up in our data in the week of September 15, 2014. We first separate out infrequent users with less than three trips during the week. We then use the same set of predictors as those used to classify users in the full month of April 2015.

The resulting clustering of users and their composition are remarkably similar to those from April 2015. Among the 6.7 million unique cards that had at least one trip in the week of September 2014, about half (3.2 million) are infrequent users; 0.8 million are classified as peak-hour commuters who have regular travel patterns; another one million cards are frequent users who travel to many destinations in both peak and off-peak hours; about one million cards travel mostly during non-peak hours on weekdays; and about 0.7 million cards mostly travel on weekends. Table

C.1 reports the characteristics of the trips by user groups. It is worth noting that the classification of user types, their characteristics, and the distribution of types among the broader user group, are all remarkably similar to those from users in April 2015 (see Table 3).

D Details of Welfare Calculations under the Current and Alternative Fare Structures

D.1 Populating Data in the week of September 15, 2014, to Full Month

We use data from September 2014 to describe the subway ridership before the fare adjustment. The data in September 2014 covers only one week between the 15^{th} and the 21^{th}. However, a key component of the current fare structure is a cumulative quantity discount that depends on expenditure during the course of the month. Therefore, we need to populate our one-week data into the full month.

Specifically, we redraw with replacement the full sample four times. We randomly keep two-sevenths of the last redrawn sample. Together with the original one-week data, the simulated sample covers a total of 30 days. In each redrawing, we set the weight of each observation (the relative probability a trip is selected) by the total number of trips observed for the user in the original data. The weighting is motivated by the observation that frequent users are more likely to have a regular travel pattern and are thus more likely to take similar trips in weeks of the month that our data do not cover. The simulated monthly ridership has 123 million trips and 1.9 billion passenger kilometers from 6.7 million unique users. On average, each card has 18.4 trips covering a distance of 289 km. Admittedly, monthly ridership populated from one-week data misses many infrequent users. Nevertheless, the populated monthly ridership is similar to the actual number. In April 2015, for which we have data for the full month, there were a total of 124 million trips from 12.4 million unique users. In a given week, there were about 28 million trips from 7 million unique users.

D.2 Algorithms to Calculate Ridership and Counterfactual Fares

Calculating ridership under the current fare structure. Starting with the populated ridership for September 2014, we calculate the counterfactual ridership under the current fare structure. This is done in the following steps.

1. First, we classify users in September 2014 according to their ridership patterns. We use the K-means clustering algorithm with the same set of predictors as we did for users in April 2015. The algorithm results in the same five groups of users with a distributional remarkably similar

to those in April 2015. Appendix C.3 reports the details. The implicit assumption imposed here is that users do not switch types due to the fare structure change.

2. We impose the demand elasticity specific to each user's type. Heterogeneous demand elasticities are reported in Figure 3, and we do not separate different types of trips within the same user type. For trips that are in the neighborhood of EBD cutoff, we impose the set of rescheduling elasticity as reported in Figure 8 Panel D. The rescheduling elasticity is -0.4 at 7 AM and 0 at 7:30 AM; we impose a linear line that connects the two points. We assume the rescheduled trips evenly land between 6:30 and 6:59 AM.

3. With demand and rescheduling elasticities at hand, we calculate the corresponding number of trips under the new fare structure for each trip in September 2014. The number of trips mostly declines because the average fare is substantially higher under the new fare structure. In addition, the EBD changes the timing for some trips.

4. We consider four different behavioral responses to cumulative quantity discounts. The oblivious, myopic, ironing, and rational users respond to the listing price, instantaneous marginal price, month-end average price, and month-end marginal price, respectively. Except for the listing price, the other three prices are functions of the monthly expenditure, which in turn is a function of the corresponding prices. We adopt an iterative algorithm to determine the ridership and the associated prices. Starting with the listing price, we calculate each user's demand for trips under the specific behavioral type. We then sum up her expenditure and update the corresponding prices. This procedure is repeated until the implied prices each user faces are consistent with her ridership throughout the month.

Calculating fare levels and ridership under alternative fare structures. We calculate the fare level under each alternative fare structure such that the aggregate revenue is the same as that under the current fare structure. Revenue under the current structure differs by how users respond to the quantity discount, so we calculate a different fare level for each behavioral type.

The algorithm follows a double-layered iterative process. We start with an initial guess of the fare level. The inner iteration obtains the ridership for each user that is consistent with the discounted prices she perceives under a given behavioral assumption. In the outer iteration, we update the fare level until the resulting revenue converges with the target.

D.3 Extrapolating Road Speed in Beijing

To calculate the marginal cost of traffic congestion (MECC, Equation 9), we need to measure a city-wide speed by hour and by day of the week. The main source of the speed measure comes from the replication data of Yang et al. (2020), who use hourly speed data in 2014 from a set of road monitoring stations on Beijing's highways and ring roads, from which they calculate speed

by the hour for each road segment.

Several adjustments are made to fit our needs. First, their data only cover weekdays. To impute speed on weekends, we use the road segment level speed from Baidu Maps, used in Gu et al. (2021b), which have hourly speed for a non-randomly selected sample of road segments for both weekdays and weekends. We calculate the average ratio between weekday and weekend speed for each hour in the same segment. The average weekend-to-weekday ratio by the hour is then used to impute the weekend speed for each hour in the road monitoring data. Second, we re-weight speed from road monitoring stations by location to generate an average speed that is relevant for subway trips. The location of a road monitoring station is denoted by its position relative to the ring roads. Beijing has five ring roads, from the second ring road (closest to the city center) to the sixth ring road.[5] The five ring roads cut Beijing into six areas ranked by the distance to the city center. We first take the simple average of speed among monitoring stations that are in the same area. We then weigh the average speed in each area by the spatial distribution of subway ridership: Consider a trip that goes from station O to station D, we calculate the track distance of the trip that falls into each of the six areas. We then sum up all trips to obtain the spatial distribution of subway trips. We calculate that about 22% of passenger-kilometers from all subway trips take place within the second ring road, 34% between the second and third ring roads, 28% between the third and fourth ring road, and 11% beyond the fourth ring road.

Finally, it is worth noting that road monitoring stations cover only highways and ring roads (which are also closed expressways). Even during congested hours, speed on those roads remains much higher than that on local roads. Indeed, the average speed recorded by the monitoring stations is about 67 km/hr. Even the lowest speed, measured during evening rush hours on roads within the second ring road, is about 52 km/hr. It is likely an overestimation of the actual average speed experienced by a traveler. Hourly speed from Baidu Maps in the sample of road segments used by Gu et al. (2021b) is about 30 km/hr on average. Everything else equal, Equation 9 suggests that a higher speed leads to a lower marginal cost of traffic congestion. Therefore, our calculations of congestion externality are likely an underestimate.

D.4 Additional Results on Ridership and Welfare under Alternative Fare Structures

Aggregate ridership and welfare under alternative fare structures with ironing and myopic consumers. Table 7 reports the aggregate ridership and welfare under alternative fare structures with rational and oblivious users. Table D.1 reports the same metrics with ironing and

[5] There is no first ring road. The first or inner ring customarily refers to the set of roads surrounding the Forbidden City. but unlike the other five ring roads, this set of roads are not express roads with dedicated ramps for access.

Table D.1: Aggregate Ridership and Welfare under Alternative Fare Structures with Ironing and Myopic Consumers

			Ironing			Myopic	
	orig. flat rate	current fare	alt. flat rate	peak/ off-peak	current fare	alt. flat rate	peak/ off-peak
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Alternative prices (yuan)							
Flat rate	2		3.97			4.00	
Listing p for a 6 km ride		3		4.50 (peak)	3		4.50 (peak)
Avg. listing p/km	0.13	0.30	0.25	0.32	0.30	0.25	0.32
Panel B: Ridership (per user per month)							
Monthly revenue (yuan)	36.8	59.3	59.3	59.5	59.5	59.5	59.5
# of trips	18.4	15.1	14.9	15.1	15.1	14.9	15.2
Total distance (km)	289	229	235	230	231	234	232
Avg. discount rate (%)	-	3.6	-	3.8	3.6	-	3.9
Panel C: Change in welfare compared with original flat rate (yuan per user per month)							
Revenue increase	-	22.5	22.5	22.5	22.5	22.5	22.5
Consumer welfare loss	-	32.0	32.8	36.2	32.4	33.2	36.8
Deadweight loss	-	9.5	10.3	13.7	9.7	10.5	14.3
Congestion externality	-	7.0	6.4	12.4	6.9	6.4	12.4

Note: This table summarizes the fares and aggregate welfare under alternative fare structures. Pre-fare-adjustment monthly ridership patterns are based on data from September 2014. Deadweight loss is the consumer's utility loss minus the operator's gain in revenue.

myopic users. Ridership and welfare outcomes under these two heuristically optimizing behavioral types are similar and lie between the polar cases in which users are rational or oblivious. With ironing or myopic users, the consumer welfare loss and the deadweight loss are the lowest under the current fare structure. A revenue-equivalent flat rate results in slightly higher welfare losses but a somewhat smaller congestion externality. Adding a peak/off-peak pricing to the current fare structure results in substantially larger welfare losses and congestion externality.

Distributional Welfare under Alternative Fare Structures. We consider welfare under alternative fare structures separately for different user types. We consider two polar behavioral types in response to the quantity discounts: oblivious and rational. Welfare with ironing and myopic users in general lie between the two polar cases and is omitted from the table in the interest of space.

It is worth noting that although we set fare levels such that the *aggregate* revenue is the same across all fare structures, revenues from different user types need not be equal. In other words, the incidence of alternative fare structures falls differently on different user types. This generates interesting distributional welfare implications on top of aggregate impacts reported in Table 7.

Table D.2 reports four measures regarding consumer welfare and overall efficiency: (1)

Changes in consumer expenditure after accounting for the discounts, which is equivalent to changes in revenue, (2) changes in consumer welfare, (3) changes in the deadweight loss, and (4) changes in the congestion externality. All measures are in yuan per user per month, and the changes are relative to the original fare structure with a 2-yuan flat rate. We report the corresponding percentage change relative to each group's average expenditure under the original 2-yuan flat rate in brackets.

For frequent commuters with a regular travel pattern (Panel A), the current fare structure performs worse than an alternative flat rate if users are oblivious to the quantity discounts. The average consumer welfare loss is equal to 90% of the average expenditure of this user group under the original 2-yuan flat rate. The consumer welfare loss is much higher than the increase in revenue. As a result, the deadweight loss is equal to 54% of the original expenditure. Under the alternative flat rate, the consumer welfare loss and deadweight loss are 79% and 42% of the original expenditure, respectively. Changes in the ridership under the current fare structure generate a congestion externality of 36% of the original expenditure, compared with 21% under the alternative flat rate.

However, If users are rational, the welfare and efficiency consequences of the two alternative structures flip. The consumer welfare loss, the deadweight loss, and the increase in congestion externality are all smaller under the current fare structure. The results are intuitive. The quantity discounts built into the current fare structure are designed to cross-subsidize frequent users. They result in a marginal price that is lower than the listing price for those who spend enough to qualify for the discounts. Rational users take advantage of the discounted prices and take more trips, which leads to lower losses in consumer welfare and a smaller congestion externality.

Compared with the current fare structure, adding a peak-hour premium always hurts regular commuters irrespective of the behavioral type. This is because regular commuters take most trips during peak hours. Loss in consumer surplus is 28% higher (95.2 versus 74.5) if users are oblivious and 45% higher (94.2 versus 65.5) if they are rational. The deadweight loss is 33% (59.8 versus 44.9) and 135% (38.8 versus 16.5) higher, respectively. Peak/Off-peak pricing also generates the largest congestion externality because it effectively reduces peak-hour subway trips; some of these trips will be taken on surface roads, while an additional vehicle on busy roads during peak hours generates a particularly large negative externality.

The welfare impacts for the frequent users with more diversified trips (Panel B) and the relative performance of alternative fare structures are similar. Compared with regular commuters, the deadweight loss as a percentage of the average monthly expenditure under the original 2-yuan flat rate is smaller because less-regular commuters have a more inelastic demand.

For less frequent users (weekday non-rushers, weekenders, and infrequent users, Panels C through E), the alternative flat rate almost always performs better than the current fare structure.

The loss in the consumer surplus, increase in the deadweight loss, and the additional congestion externality are mostly lower. It is true under both behavioral assumptions, particularly when users are rational. This is because less frequent users rarely qualify for the quantity discounts, and they cross-subsidize the frequent users under the current fare structure. The magnitude of the cross-subsidy is larger if users are rational, when the frequent users take full advantage of the discounts, effectively paying a much lower price than less-frequent users.

Consumer welfare losses under the fare structure with peak-hour premiums are generally smaller for less frequent users. This is because those users mostly take trips during off-peak hours. Interestingly, even for those users, peak/off-peak pricing still generates a larger congestion externality than the other two fare structures. This reflects the extremely steep function of the marginal congestion externality with regard to road density, which is much higher during rush hours.

Temporal Ridership Patterns and user composition under alternative fare structures. Figure D.1 shows the temporal distributions of subway trips under different fare structures. Ridership in the pre-fare-adjustment period is represented by the ridership in September 2014 populated into the full month (see Appendix D.1). The graphs show the composition of trips from different user types, with each represented by a different color.

Panel A shows the ridership under the original 2-yuan flat rate. There are two clear peaks on weekdays, one between 7 and 9 AM and the other between 5 and 7 PM. During peak hours, the ridership is between two and three times the daily average. Regular and less-regular commuters account for the bulk of peak-hour trips. Ridership during weekends is lower, and there is less variation throughout the day. Weekenders and less-regular commuters account for the majority of weekend subway trips.

Panels B through D report ridership distributions under three revenue-equivalent fare structures. The ridership, fare, and revenue depend on how users respond to the quantity discounts, so in each panel, we report two graphs with oblivious (on the left) and rational (on the right) users.

Aggregate and distributional ridership and welfare under alternative fare structures with different user behaviors are reported in Table 7 and Appendix Table D.2. Here, we focus on how different fare structures can pare peak demand during rush hours, causing crowdedness in the subway system.

First, compared with the original 2-yuan flat rate, peak demand under any of the three post-adjustment fare structures is substantially reduced due to higher fare levels. Among the three alternative fare structures, the one with peak/off-peak pricing is the most effective in cutting peak load, regardless of behavioral assumptions. Peaking/Off-peak pricing mostly achieves that by reducing regular and less regular commuters' trips. Few trips are rescheduled to less busy hours in response to the price difference between peak and off-peak hours: The numbers of trips

between 6-7 AM, 9-10 AM, 4-5 PM, and 7-8 PM, respectively, are similar between Panel B and Panel D.

Given the fare structure, there are larger reductions in peak-hour demand if users are oblivious. This is because predominant users during peak hours are frequent users who qualify for discounts. If they are rational, they take advantage of the discounts and reduce their trips to a lesser extent. Although we do not quantify the welfare implications due to crowded subway cars, graphs in Figure D.1 illustrate the tradeoff between consumer welfare, congestion externality, and negative externality from crowdedness.

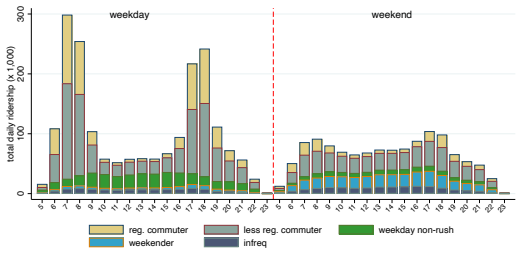
Table D.2: Ridership and Welfare under Alternative Fare Structures: Distributional Impacts

	Oblivious			Rational		
	current fare	alt. flat rate	peak/ offpeak	current fare	alt. flat rate	peak/ offpeak
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Regular commuters (avg. $EXP_0 = 82.5$ yuan)						
Change in expenditure (revenue)	29.6	30.7	35.3	49.0	37.8	56.1
[% of EXP_0]	[35.9]	[37.2]	[42.8]	[59.4]	[45.8]	[68.0]
Loss of consumer surplus	74.5	65.0	95.2	65.5	81.1	94.9
	[90.3]	[78.8]	[115.3]	[79.4]	[98.3]	[115.0]
Deadweight loss	44.9	34.3	59.8	16.5	43.3	38.8
	[54.4]	[41.6]	[72.5]	[20.0]	[52.5]	[47.0]
Congestion externality	29.5	21.4	47.5	15.1	24.8	33.4
	[35.8]	[25.9]	[57.6]	[18.3]	[30.1]	[40.5]
Panel B: Less-regular commuters (avg. $EXP_0 = 109.9$ yuan)						
Change in expenditure (revenue)	52.7	67.4	60.8	69.7	84.7	77.0
[% of EXP_0]	[48.0]	[61.3]	[55.4]	[63.4]	[77.1]	[70.1]
Loss of consumer surplus	84.2	90.4	98.0	76.2	224.2	97.2
	[76.6]	[82.3]	[89.1]	[69.3]	[204.0]	[88.4]
Deadweight loss	31.5	23.0	37.4	6.5	29.4	20.5
	[28.7]	[20.9]	[34.0]	[5.9]	[26.8]	[18.6]
Congestion externality	21.3	14.9	38.9	8.7	17.5	26.4
	[19.4]	[13.6]	[35.4]	[7.9]	[15.9]	[20.0]
Panel C: Weekday non-rushers (avg. $EXP_0 = 34.8$ yuan)						
Change in expenditure (revenue)	29.1	22.9	20.0	31.1	28.9	24.4
[% of EXP_0]	[83.6]	[65.8]	[67.9]	[89.4]	[83.0]	[70.2]
Loss of consumer surplus	37.8	28.4	32.9	37	35.9	32.1
	[108.6]	[81.6]	[94.5]	[106.3]	[103.2]	[92.1]
Deadweight loss	8.7	5.5	9.3	5.9	7.0	7.7
	[25.0]	[15.8]	[26.8]	[17.0]	[20.1]	[22.1]
Congestion externality	5.5	3.8	7.3	4.2	4.5	6.2
	[15.8]	[10.9]	[21.1]	[12.1]	[12.9]	[17.7]
Panel D: Weekenders (avg. $EXP_0 = 26.2$ yuan)						
Change in expenditure (revenue)	21.1	15.4	15.6	15.6	19.3	15.7
[% of EXP_0]	[80.5]	[58.8]	[59.4]	[83.6]	[73.7]	[59.8]
Loss of consumer surplus	29.9	21.0	23.3	29.6	26.4	22.6
	[114.1]	[80.2]	[88.8]	[113.0]	[100.8]	[86.3]
Deadweight loss	8.8	5.6	7.8	7.7	7.1	7.0
	[33.6]	[21.4]	[29.7]	[29.4]	[27.1]	[26.8]
Congestion externality	5.1	3.6	5.5	4.7	4.2	5.1
	[19.5]	[13.7]	[20.9]	[17.9]	[16.0]	[19.3]
Panel E: Infrequent users (avg. $EXP_0 = 5.3$ yuan)						
Change in expenditure (revenue)	5.2	3.5	4.3	5.2	4.4	4.2
[% of EXP_0]	[98.1]	[66.0]	[80.9]	[98.1]	[83.0]	[78.9]
Loss of consumer surplus	6.8	4.6	6.2	6.8	5.8	6.0
	[128.3]	[86.8]	[116.2]	[128.3]	[109.4]	[113.2]
Deadweight loss	1.6	1.1	1.8	1.6	1.4	1.8
	[30.2]	[20.8]	[34.5]	[30.2]	[26.4]	[33.6]
Congestion externality	0.9	0.6	1.2	0.9	0.7	1.2
	[17.0]	[11.3]	[22.5]	[17.0]	[13.2]	[22.1]

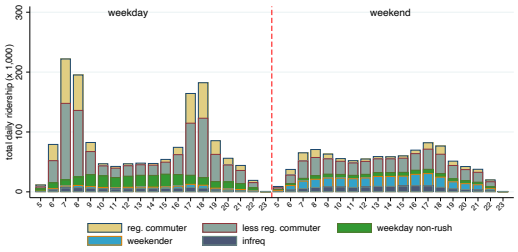
Note: This table summarizes ridership and associated welfare under alternative fare structures for different user types. Counterfactual fare levels are calculated such that the aggregate revenue is the same as under the current fare structure with the specific behavioral responses to the quantity discounts. Two behavioral responses are considered: oblivious and rational. Numbers are in yuan per user per card. Numbers in brackets are percentage points relative to the user type's average expenditure under the original 2-yuan flat rate.

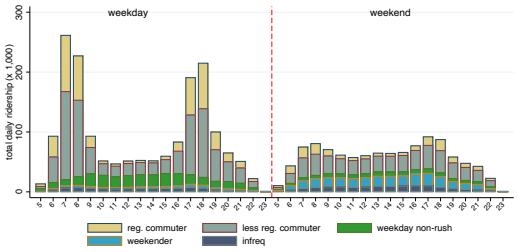
Figure D.1: Compositional and Temporal Travel Patterns under Alternative Fare Structures

Panel A: Under 2-yuan Flat Rate

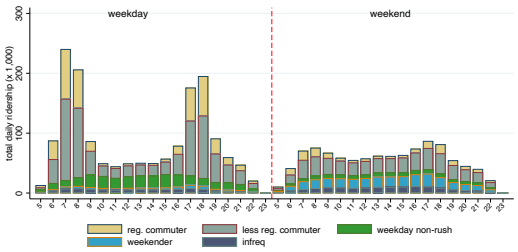


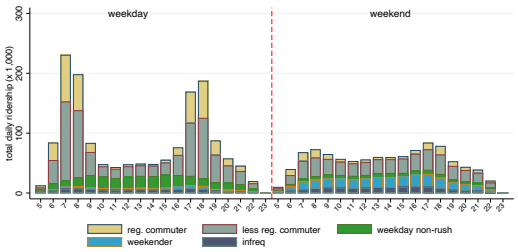
Panel B: Under the Current Fare Structure



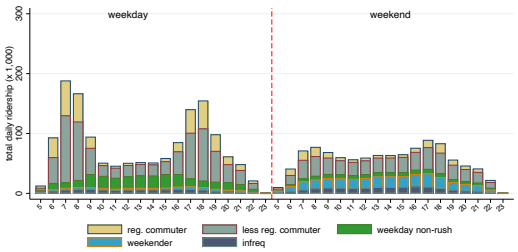


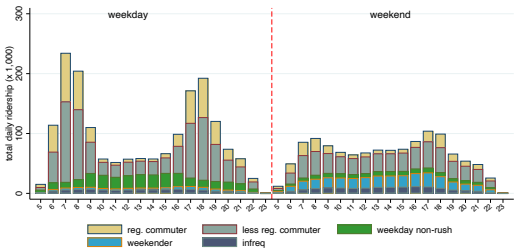
Panel C: Under the Alternative Flat Rate





Panel D: Under the Alternative Peak/Off-peak Pricing





Oblivious Rational

Note: 10% random sample of users from simulated full-month data from the September 2014 data. Graphs on the left target the aggregate revenue in the current fare structure with oblivious users. Graphs on the right assume rational users.