Fare Structure and the Demand for Public Transit

Yizhen Gu Qu Tang Yacan Wang Ben Zou*

March 2024

Abstract

We study fare structure design and public transit use. Leveraging a fare rise in the Beijing Subway that replaced a flat rate with one that varies by distance, time, and offers quantity discounts, we find inelastic demand, inflexible travel schedules, and no responses to discounts. The new fare structure generates welfare losses due to lower consumer surplus and higher surface road congestion that far exceed revenue gains. The new fare is less efficient than a revenue-equivalent flat rate when users are unresponsive to discounts, while fare structures featuring a peak-hour premium are worse due to externality on road congestion.

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.

1 Introduction

Public transit plays an indispensable role in urban life. 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 typically focus on the overall efficiency under a single-dimension fare level and with 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 design of the fare structure is reflected in the distinctly different price schedules 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.[1]

This paper provides the first comprehensive study on how designs of the fare structure affect user behavior and determine the efficiency, externality, and welfare impacts of the transit system. The empirical setting is the Beijing Subway, which is 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) the 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 cumulative expenditure. The main data are the universe of trip-level records captured by an electronic payment system in periods both before and after the fare structure change.

The fare adjustment provides useful variations along several dimensions, which allow us to estimate three sets of key parameters capturing user demand. 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 take advantage of price differences over time. We test how consumers respond to kinks in the budget constraint generated by the

[1] For the pricing of London Tube, see https://tfl.gov.uk/fares/find-fares/tube-and-rail-fares. For the pricing of the NYC Subway, see https://new.mta.info/fares. Last accessed in March 2024.

cumulative quantity discounts. Those parameters capture general user behavior and have broad implications for other major public transits.

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

The new price as a step function of distance creates sharp discontinuities in the fare between trips that vary only slightly in distance. We adopt a regression discontinuity (RD) design to estimate the demand elasticity. Users may strategically shorten a trip in order to bunch below the distance cutoff and save cost. We design several approaches to account for strategic bunching behavior. Our preferred demand elasticity is around -0.36. .

While the data do not have information on users' demographics, trips from the same user can be linked via an anonymized ID, which allows us to classify users into groups by ridership patterns. Commuters who have regular temporal and location patterns have the largest demand elasticity, at around -0.55. Within each user type, the elasticity is similar across trips taken at different times and possibly for different purposes. This suggests that users do not systematically substitute between different kinds of trips.

The quantity discounts based on cumulative expenditure create a non-trivial optimization problem for the user. We distinguish four different behavioral responses 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 monthly average price.[2] A consumer is "myopic" if she makes decisions based on the *current* marginal price she currently qualifies for. Finally, we call a consumer "oblivious" if she only responds to the listing price.

The discounts create non-convex kinks in the monthly budget, which, under some general assumptions, would result in a "bowl" in the distribution of monthly subway expenditures around each kink (Saez, 2010; Kleven, 2016). We find no evidence for any non-smoothness 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 a statistical mixture model embedded in our baseline RD framework. We find that users predominantly respond only to the listing price. This is true even among frequent users with a regular travel pattern, for whom a quasi-optimal response to the non-linear budget generates substantial welfare gains.

The EBD creates a trough in ridership right after the time cutoff as users move their trips

[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.

earlier to qualify for the discount. However, the trough is neither deep nor wide. We find that while a ten percent EBD moves about 4% of the trips that are initially planned to depart right after the time cutoff, the effect quickly reduces to zero for those initially planned to depart 15 minutes later. Overall, the 30% EBD moves less than 1% of the rush-hour trips to off-peak hours.

Empirical results are used to evaluate the welfare impacts of the fare structure change. 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 revenue gains. 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 a negative externality on surface roads. The cost of the resulting congestion externality is equivalent to 41% of the revenue gains.

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

Finally, we evaluate alternative fare structures. While in theory, the new fare structure generates a good balance between 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 fares poorly due to the relatively high demand elasticity among rush-hour commuters and the low rescheduling elasticity. It substantially reduces high-value commuting trips by diverting them to surface roads, which generates large congestion externalities.

Contributions to the Literature

A well-functioning public transit system is essential to making 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, compared with the voluminous literature on road pricing since Vickrey (1963), there are relatively few studies on the pricing and demand for public transit. This is particularly surprising given public transit makes up a substantial share of urban travels across the world, and cities, especially those in developing countries, have been investing heavily in building an extensive transit. This paper makes several contributions to this literature.

Existing studies typically focus on a single component of the transit price, most on the level

of the fare and aims at estimating the demand elasticity (see Cervero, 1990; Small and Verhoef, 2007, for reviews). Instead, we leverage a comprehensive, multi-faceted fare structure change to uncover a set of key demand parameters and study how they interact. The set of parameters jointly estimated from the same empirical setting is internally consistent. Together, they speak to the design of the fare structure and its welfare implications.[3]

This paper contributes to the credible identification of those key parameters by using rich trip-level data and adopting design-based identification strategies tailored to the policy experiment.[4] Largely consistent with the existing literature, we find that the demand for public transit is inelastic, and travel schedules are largely inflexible.[5] Our estimates are also among the few in the literature that come from a developing country setting. While it is intuitive to expect less-than-optimal user behavior in the setting of demand for public transit given the often-abstruse fare structure and limited user attention, much of the existing literature assumes users to be fully rational. We present convincing evidence that transit users are inattentive to the nonlinear budget constraint and highlight the importance of understanding user behavior in the design of optimal fare structure.[6]

The findings in this paper have broad implications for the design of multi-dimensional pricing schedules in other markets where consumer responses may be less than fully rational. The pricing of many regulated natural monopolies, such as electricity and water, share similarities with that of public transit. The price structures of such goods are often designed to smooth demand over time (e.g., Jessoe and Rapson, 2014), account for externalities (e.g., Reiss and White, 2005),

[3]In an influential study, Parry and Small (2009) study the optimal pricing of major transit systems in the United States and the United Kingdom. They focus on the level of optimal subsidy and the aggregate welfare impacts, and use parameter values borrowed from other studies or set at rule-of-thumb values. In contrast, key parameters in this paper are causally identified internally. Using trip-level data, this paper speaks to the distributional impacts, which have important policy implications.

[4]Digital payment methods in transit systems have produced big data with rich information, which opened up research opportunities (Pelletier et al., 2011). However, such user-generated data typically lack important demographic information, which is essential for the study of heterogeneous effects and distributional welfare impacts. This paper illustrates how to infer user types using machine learning. Similar algorithms have been used in other settings with user-generated data, such as classifying Uber users (Chen et al., 2019) and imputing household income (Cahana et al., 2022).

[5]Our estimate of the demand elasticity fits well in the distribution of estimates found in the United States and other developed countries (reviewed by Cervero, 1990; Litman, 2004; Holmgren, 2007). Using aggregate ridership data in Mexico, Davis (2021) also finds a similar elasticity. Using trip-level data and natural field experiments, Hahn et al. (2023) find a somewhat larger demand elasticity, though still smaller than one in magnitude. Using temporal differences in the fare, many existing studies find that schedules for commute trips are inflexible. Kreindler (2020) conducts a field experiment with motorists in Bangalore and finds that only a small fraction of peak-hour commutes are responsive to off-peak discounts. Hahn et al. (2023), Yang and Long Lim (2018), Ma et al. (2020) find small but meaningful rescheduling elasticities among transit users in San Francisco, Singapore, and Hong Kong. Ma et al. (2020) present additional evidence that much of the initial responses to off-peak discounts reverted after a few months.

[6]Larcom et al. (2017) document interesting and convincing evidence of suboptimal user choices in transit. They study strikes in the London Underground that forced passengers to experiment with new routes. They find that many stuck to the new routes after the strikes were over, which implies suboptimal route choices before the strikes.

and consider distributional impacts (e.g., Borenstein, 2012). There is a 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 nonlinear pricing with limited salience is unable to induce desirable consumer responses. Different behavioral responses generate substantially different aggregate and distributional welfare impacts, and imply distinctly different optimal fare structures.

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 evidence on user responses. 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 the second-largest city in terms of 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).

The city has invested heavily in the subway network in the past two decades. In 2001, the system had two lines and about 40 stations. By 2019, the system consisted of 25 lines, over 400 stations, and an annual ridership of 3.8 billion. Subway has a particular advantage in long-distance commuting trips due to its fast speed and reliability. 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 buses and bikes.

Government-owned companies operate the Beijing Subway.[7] Its fare is determined by the

[7]Beijing's public transit also includes an extensive bus system consisting of 1,158 bus lines and a fleet of 25,000 vehicles. 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

municipal government and is heavily subsidized. Before the fare adjustment in 2014, The subway had a flat-rate fare of 2 yuan (approximately 0.3 US dollars) 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 The Fare Structure Adjustment

In December 2014, the Beijing Subway overhauled its fare structure. The simple flat rate was replaced by a more complex fare structure that contains 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 an electronic smartcard for payment. A user starts each calendar month by paying the listing price but qualifies for a 20% discount for future trips once her month-to-date cumulative expenditure reaches 100 yuan. The discount rate rises to 50% once the out-of-pocket expenditure reaches 150 yuan. Discounts are capped at the out-of-pocket expenditure of 400 yuan, after which the user receives no discount for trips in the remainder of the month. Third, starting in December 2015, An early-bird discount (EBD) was applied in 16 stations that are often crowded 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 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.[8] 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 temporal distribution of trips and reduce crowdedness in the system during peak hours. The EBD was intended to achieve this goal by creating a price difference over time. Its success depends crucially on the flexibility of

bus remained a far more affordable option. This paper focuses on the impacts of subway fare adjustment on demand for subway trips and does not consider interactions between the subway and the bus.

[8]See, for example, the following interview (in Chinese) an official in the municipal Department of Public Transportation. http://politics.people.com.cn/n/2013/1219/c1001-23885064.html (last accessed on September 2, 2023).

users' travel schedules, captured by the "rescheduling elasticity" we estimate below. The third goal was to maintain relatively fair distributional impacts despite the substantial fare hike. While frequent subway users and those with long trips stand to pay a lot more, the cumulative quantity discounts aim to alleviate their financial burden. The discounts create kinks in the monthly budget. Whether the goal about welfare incidence 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). Many transit systems share the pricing features with the Beijing Subway's new fare structure. But a complex, multi-faceted fare structure is not universal. Some transit authorities are wary of sub-optimal user responses to a complex fare structure and opt for a simple pricing schedule.[9]

Table 1: Summary of Subway Fare Structures

Before Dec. 28, 2014				
flat rate of 2 yuan for all trips				
After Dec. 28, 2014				
		cumulative quantity discounts[†]		
	distance-based	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 before 7 AM*		30%		

[†] The cumulative quantity discounts rest at the start of each calendar month.
* The 16 stations are on the Batong Line and the Changping Line.

[9]For example, the pricing of the London Tube shares many similarities with Beijing's new fare structure, while the New York Subway uses a flat rate.

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 only card users qualify for discounts. By comparing the official statistics with our data, we estimate that between 90% and 95% of the subway trips were paid for by a smartcard around the time of the fare adjustment. The smartcard captures the entry and exit times and stations 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 captured by smartcards and took place in the week between September 15 and September 21, 2014 (before the fare adjustment), the full month of April 2015 (after the fare adjustment), 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 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 262 stations in the September 2014 data. About 67,000 origin-destination station pairs have non-zero ridership.

3 User Responses to the Fare Structure Adjustment

Variations created by the fare structure change allow for the estimation of key parameters that capture user demand for public transit. In this section, we first estimate the demand elasticity for subway trips by exploiting price discontinuities in distance. While our data do not have users' demographic information, we classify users by travel patterns and estimate heterogeneity in demand elasticity by user types. We then investigate behavioral responses to the cumulative quantity discounts. Finally, we estimate the rescheduling elasticity by exploiting the temporal price discontinuity created by the EBD.

3.1 Demand Elasticity

3.1.1 Step-wise 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 origin-destination (OD) 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 in ridership for trips in distance bins that are immediate

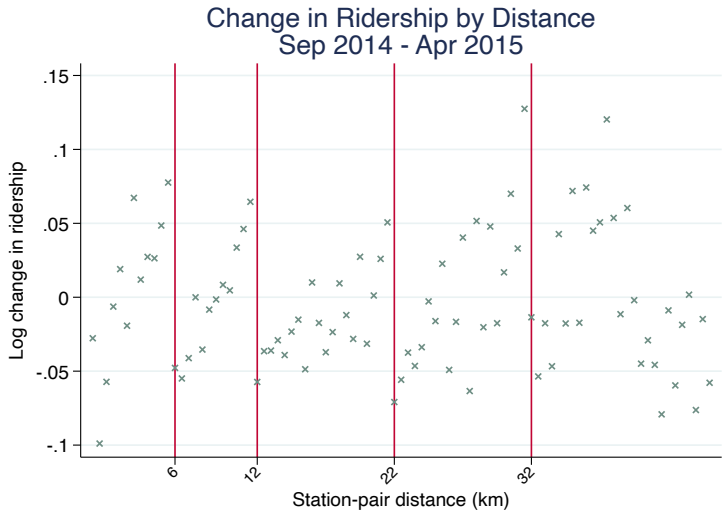
to the right of the cutoffs.

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

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

Each observation is an OD pair with non-zero ridership in September 2014. OD pairs are assigned to cutoffs in non-overlapping windows. The dependent variable is the log change in ridership between 2014 and 2015. $f(dist_{od})$ 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 trips. $\Phi_{o,d}$ is a vector of control variables that account for the heterogeneity in ridership changes. The inclusion of these variables aims at improving the precision of the estimation and is not crucial for identification. In the baseline, $\Phi_{o,d}$ includes the origin and the destination fixed effects. Each observation is weighted by the OD pair's ridership in September 2014. Therefore, e^{RD} can be interpreted as the demand elasticity for a typical trip.

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 the fare.

The results of estimating Equation 1 at the first four distance cutoffs are reported in Columns 1 through 4 of Table 2. The estimates are precise, and the point estimates are bounded in a tight range between -0.35 and -0.45. The magnitude of the elasticity does not appear to be a

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: Each observation is an OD pair. The dependent variable is the log change in ridership between September 2014 and April 2015. Columns 1 through 4 report estimations of Equation 1 at each of the first four distance cutoffs. In Columns 5 and 6, all distance cutoffs are included. In Column 6, OD pairs that are immediate to a distance cutoff are excluded. Robust standard errors are in parentheses.

monotone function of distance.[10] 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 reports a demand elasticity of -0.39.

We conduct common checks to test the validity of the RD design (Cattaneo et al., 2019) and 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 reports the details of the validity tests and robustness checks.

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 trip by entering the station further downstream or exiting before arriving at the ideal destination station. Such bunching behavior represents little change in the actual use of the subway in terms of passenger-kilometers but would contribute to a larger difference in the number of trips on different sides of the cutoff, as is captured by the RD estimation.

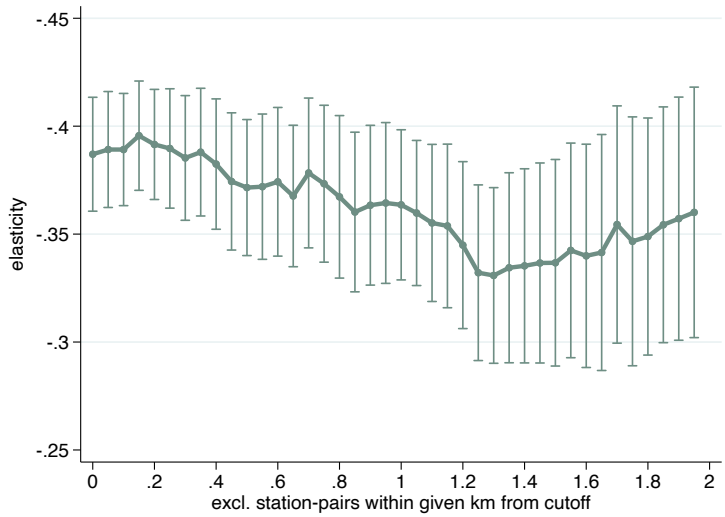
We adopt a "donut-hole" RD design (Cattaneo et al., 2019) to estimate the demand elasticity that is not contaminated by strategic bunching. Adjacent OD pairs immediately around the cutoffs are excluded from the regression.[11] For the donut-hole RD estimation to be valid, users are

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

[11] Consider an OD pair (O, D), let $D+1$ denote the next station after D. If (O, D) is below a cutoff while $(O, D+1)$

assumed to have no incentive to bunch beyond immediate OD pairs around the distance cutoff. The estimated demand elasticity is -0.36 (Table 2 Column 6). Consistent with the existence of strategic bunching, it is smaller in magnitude than that from the whole sample (-0.39, Column 5). Still, the difference is small, suggesting that strategic bunching is limited. Column 6 is our preferred baseline estimate.

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 have 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 assumption that strategic bunching is limited to immediate OD pairs around a cutoff is intuitive. Bunching incurs time and pecuniary costs: the passenger may need to walk or bike for a longer distance, take a detour, or ride a bus to make up for the shortened subway trip. Given the subway's clear advantage in speed and reliability over those other modes, substantially cutting the subway trip to save a mere one yuan is unlikely to be worthwhile. In addition, because price is a step function of distance, users cannot always save more by bunching beyond the immediate station pairs.

To test this assumption, we re-estimate Equation 1 by varying the size of the donut hole. As the window around the cutoff 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 eventually stabilize as essentially 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

is above, both pairs are dropped from the sample. About 20% of the OD pairs are dropped due to this restriction.

hole is about 1.2 km wide in radius (2.4 km in diameter), and largely stabilizes at around -0.35 afterward. The median distance between adjacent stations is about one kilometer. A 2.4-km-wide 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 -0.3 and a little over -0.4.

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. With that estimate, we get a demand elasticity adjusted for bunching 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 using linked users is around -0.27. We show that user-level estimates are immune to the bunching behavior but are an underestimate of the true elasticity.

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, 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 the demand elasticity reported in Table 2 governs how the aggregate ridership would change as a result of the fare rise, it does not speak to *who* are affected and in what ways.

In lieu of demographic information, we classify user (i.e., card) types according to their travel patterns.[12] That task is achieved by adopting a $K-$means clustering algorithm. With a predetermined number of clusters (the K) and a set of chosen predictors, the algorithm groups users into clusters to minimize within-cluster distances. The appropriate value of K can be cross-validated by inspecting the patterns of each cluster. We stop increasing the number of clusters when allowing for a larger K does not lead to a new cluster with a distinct pattern.

[12]In Appendix E, we present novel data of commuting flows at fine graphic levels that come with *imputed* education levels of commuters. The data are derived from information collected by location-enabled smartphones. We present a method that maps the skill composition in the smartphone-based commuting flow data to that of subway users in any station pair. We find that skilled subway riders have a slightly lower demand elasticity, and the welfare incidence of the fare structure change falls roughly equally on skilled and unskilled users.

We apply this algorithm to classify the 12 million users in the April 2015 data. 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 occurred 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* concentration 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 a user took 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. We name each with its most salient pattern: weekenders (1.98 million), weekday off-peak users (2.74 million); rush-hour commuters (1.3 million), and all-purpose users (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 off-peak users have an average of nine subway trips, and over half of those are during off-peak hours on weekdays.

Both rush-hour commuters and all-purpose users are frequent subway riders. The main difference between the two groups is that rush-hour 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). All-purpose users also use the 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.

Appendix Figure C.1 shows the distribution of trips by user type and time. The average ridership on a weekday is 50% higher than on a weekend. Rush-hour commuters and all-purpose users account for 65% of the weekday ridership and over 80% of the peak-hour system load. The same algorithm can be applied to users in the September 2014 data.

3.2.2 Demand Elasticity by User Type

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 presumably a commute if a regular commuter rides the subway during the peak hour and the

Table 3: User Classification and Characteristics

	infreq. users	weekenders	weekday off-peak users	rush-hour commuters	all-purpose users
	(1)	(2)	(3)	(4)	(5)
# of users (million)	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. Standard deviations are in parentheses. 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.

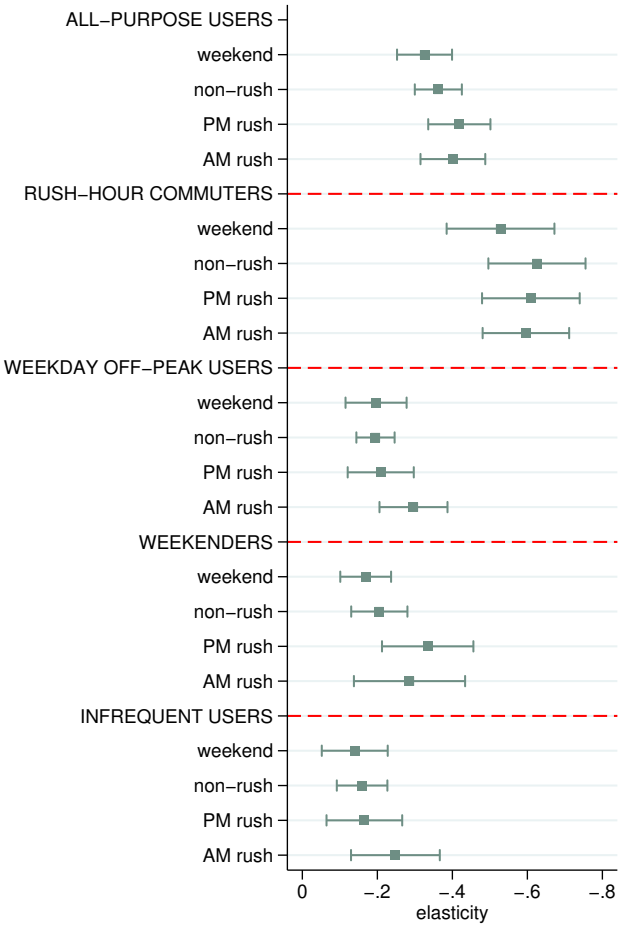
trip is between her usual OD pairs.

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 od by user-type k during time h in April 2015 and the ridership in the same OD pair during the same hour by *all* users in September 2014.[13]

Figure 3 summarizes the heterogeneity in demand elasticity. All coefficients are precisely estimated. Rush-hour commuters have the largest elasticity of about -0.55. One may have expected a smaller elasticity for this group because it is costly to replace regular commuting trips. On the

[13]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, and because data from September 2014 cover only one week. Nevertheless, Appendix C shows that classifying users from September 2014 using the same set of predictors yields strikingly similar types of users and their distribution. Defining $\Delta\ln(N_{od}^{kh})$ based on the same type of users results in quantitatively similar results.

Figure 3: Heterogeneous Effect by User 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.

other hand, they also have a stronger incentive to replace subway rides with other means because the potential savings are large. Their regular, predictable travel patterns may also imply a lower switching cost. The difference in switching cost may explain the slightly smaller (in magnitude) elasticity of -0.4 among the all-purpose users. Less-frequent users have a more inelastic demand. The elasticity for the other three groups is around -0.2.

The larger demand elasticity among commuters has implications for the distributional impacts of the fare adjustment and externality on road congestion. In the short run, changes in work or home locations are presumably second-order. Workers need to fulfill their commutes using other modes of transportation; this is likely to generate large congestion externalities on surface roads.

Another salient pattern from Figure 3 is that while the demand elasticity varies by user type, within the same type, it does not substantially differ by the type of trip. Rush-hour commuters have a higher demand elasticity. That elasticity applies 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 the null hypothesis that demand elasticities within the same user type but across different times are statistically identical. This suggests that users do not cross-substitute between trips of different types and for different purposes. For example, in response to the fare rise, it is possible that commuters would disproportionately cut non-commuting trips to save money while preserving most commuting trips. If that is the case, we would expect a larger demand elasticity for the former and a smaller elasticity for the latter. We find no empirical support for this.

3.3 Behavioral Responses to Cumulative Quantity Discounts

3.3.1 Behavioral Types

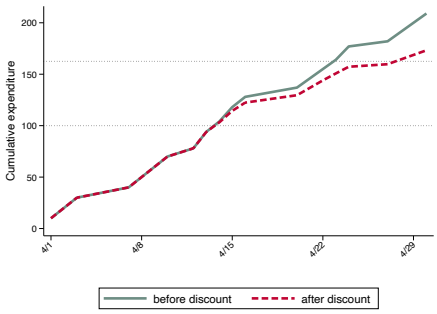
The analyses so far have exploited discontinuities in the *listing* price created by the new fare structure. However, the quantity discounts create kinks in the monthly budget and make the *out-of-pocket* price potentially different from the listing price. Understanding how users respond to the new fare structure 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 each day 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. which she achieves in about two weeks. 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 took trips worth 209 yuan, for which she paid 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.

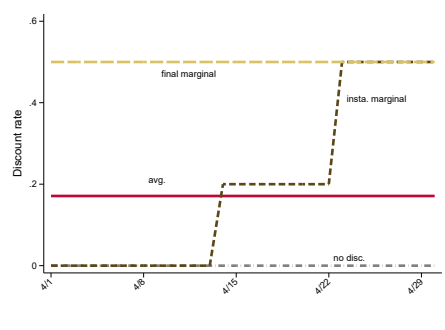
Panel B shows the paths of different discount rates to which the user may respond. A fully forward-looking and *rational* consumer responds to the *end-of-month marginal* discount rate (which we refer to as the *final marginal* discount), which is 50%. It is important to note that it is the marginal discount rate the consumer faces regardless of the date she plans to take an additional trip. Suppose a user has all the trips planned for the month, and then she ponders whether she should take another trip at the beginning of the month. Although she would not get any discount from that trip itself, it would "bump" a trip later in the month into the 50%-off

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

Panel A: Subway Expenditure by Date



Panel B: Four Perceived Discount Rates



Note: The graphs illustrate the expenditure and associated prices using the example of one user observed in April 2015. Panel A shows the 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.

region.

Being rational requires a user to fully predict her demand for subway trips for the entire month and react optimally to the marginal price. In reality, users may respond heuristically. One quasi-rational type is responding to the monthly-*average* discount rate. Responding to the average price instead of the marginal price is a common behavioral bias (Liebman and Zeckhauser, 2004). In the public finance literature, such an agent is called an *ironer* as if she mentally "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 discounts she would receive for future trips. As is shown in Figure 4, Panel B, the instantaneous marginal discount rate could change over the course of the calendar month.

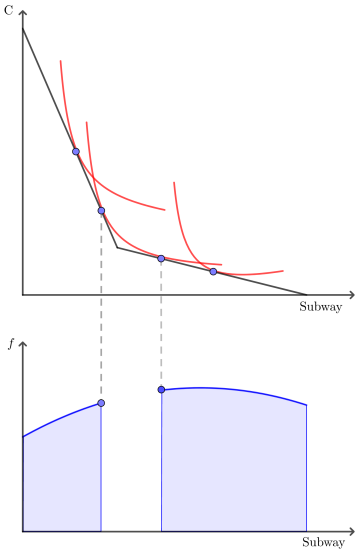
Finally, we say the user is *oblivious* to discounts if she responds only to the listing price.

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 the agent's 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.

3.3.2 A Test of Rationality

We start with a test of consumers being rational. The test leverages the fact that discounts create kinks in the monthly budget. With some additional assumptions, the pattern of user distribution around the kinks indicates the extent to which users are optimally responding to the monthly budget (Saez, 2010; Kleven, 2016).[14]

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 subway trips. The y-axis of the top graph, C, indicates the consumption of the numeraire good. The y-axis of the bottom graph, f indicates the density of users with the corresponding pre-discount monthly expenditure on subway trips.

Figure 5 illustrates the intuition of the test. We start with a simple case where users have the same demand elasticity but differ in their idiosyncratic preferences for subway trips. In the graphical illustration, it means that indifference curves from different users have the same shape but differ in the 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 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 level of subway consumption that falls in 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

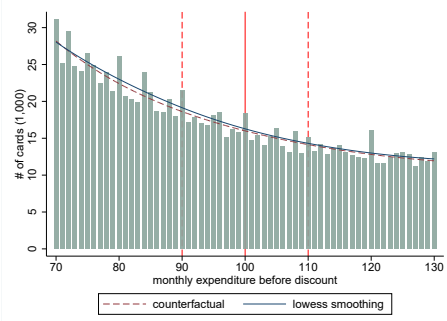
[14]The discount schedule creates two non-convex kinks (at 100 yuan and 162.5 yuan before discounts) and one convex kink (at 662.5 yuan before discounts). We focus on the first two non-convex kinks because there are few users around the neighborhoods of the third kink.

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 there will be no hole in the distribution. In contrast, if general consumption and subway trips are close substitutes, the indifference curve is close to a linear line, and the demand elasticity is large. The non-convex kink will lead to a large hole in the distribution.

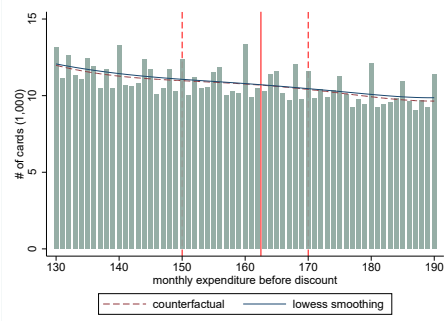
In reality, we may not see a strict hole in the distribution around a non-convex kink. There are two main reasons for that. First, users may be heterogeneous in the *shape* of their indifference curves, corresponding to differences in the demand elasticity. In this case, the hole becomes a "bowl." The kink point remains weakly dominated, so the bowl touches zero as long as the number of consumers with a Leontief preference is negligible. The width of the bowl at its mouth still allows for the estimation of the *average* demand elasticity. Second, users 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: Density Distribution around Non-convex Kinks

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 discount schedule creates three kinks in the budget line. Here, we show the distribution of pre-discount expenditures 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. In Panel B, the joint test of $\hat{\gamma}_j$ for j between the neighborhood has a $p-$value of 0.5, and the summation of those $\hat{\gamma}_j$'s is 3.24, while the average N_j within this range is 10.88.

Figure 6 shows the empirical distribution of users by their pre-discount monthly expenditure on subway trips. Each graph plots a 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 scatterplot 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) in 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 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 Leontief 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.[15] The sharp responses to differences in the listing prices also suggest that it is unlikely users face exorbitantly high adjustment costs.

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

3.3.3 A Simple Test of Myopia

If users do not respond to the final marginal price, do they respond to the *instantaneous* marginal price? We devise a simple test of 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 OD pairs likely receive *increasingly* larger discounts compared with those in below-the-cutoff pairs. There will also be distance discontinuities in the other three prices, 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 only discontinuities instantaneous marginal price evolve over time. 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 final 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 corresponding price type.

Equation 3 is separately run for each Monday through Thursday in April 2015.[16] $\rho_t^{\text{price type}}$ captures the discontinuity in the specific price around distance cutoffs. ρ_t^{listing} is equal to one by definition. We expect $0 < \rho_t < 1$ under other price types because trips in the above-the-cutoff OD pairs qualify for larger discounts. 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.

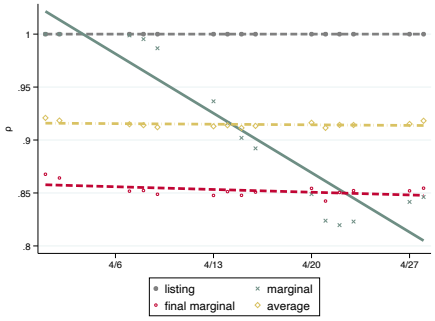
Panel A of Figure 7 plots $\rho_t^{\text{price type}}$. Except for ρ_t^{listing}, they are all below one. ρ_t^{mar} and ρ_t^{avg} are largely constant over t. Mild fluctuations around the flat fitted lines reflect small day-to-day

[16]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. 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 have substantive leisure trips and have a user composition that is different from those between Monday and Thursday.

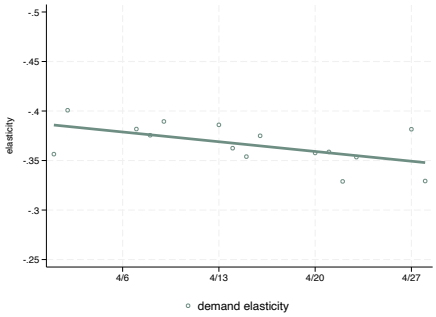
differences in the composition of users. In contrast, $\rho_t^{\text{inst mar}}$ declines over time. In fact, the path of $\rho_t^{\text{inst mar}}$ follows a flipped $S-$shape. It is close to one at the beginning of the month when no one is yet qualified for any discount. Then those who take slightly 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. Note that by definition, the instantaneous marginal price on the last day of the month is equal to the month-end final marginal price. The graph confirms that $\rho_t^{\text{inst mar}}$ and ρ_t^{mar} converge by the end of the month.

Figure 7: A Test of 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 for each day between Monday and Thursday in the month of April 2015. The regression equation is described in Equation 3. Coefficients associated with each estimation are plotted, and lines in the corresponding color are linear fits of those coefficients. The linear fitted line for $\hat{\rho}_t^{\text{inst mar}}$ 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 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_t^{listing} is the demand elasticity associated with the listing price. It has a t subscript because Equation 4 is estimated separately for each date t.

We assume the *true* demand elasticity is 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 over t because the discontinuity in the listing price increasingly *over*states the discontinuity in the instantaneous marginal 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 the 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%.[17]

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. In reality, some users may be more sophisticated and know how to behave optimally. This subsection aims at estimating the composition of behavioral types in the user population. We run an OD pair RD regression where all four prices are included.

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

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. $\ln(p_{od,t}^{\text{price type}})$ is defined as in Equation 3. $f(X_{od})$ 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 the model in Column 6 of Table 2. Standard errors are clustered at the OD-pair level.

The variation used to identify each price separately 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 in Figure 7 Panel A, the discontinuity is largest in the listing price and is typically the smallest in the final marginal price. Second, for all price types except for 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.

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

Equation 5 is more than a horse-race model. 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 as:

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

e is the demand elasticity. γ's represent the composition of users types and add up to one. 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 demand elasticity, are also more likely to be rational. To alleviate this concern, we also estimate the model separately for sub-groups. 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).[18]

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 final marginal price and the average price to capture imperfect optimization. Equation 6 becomes:

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

$g_1(.)$ 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 value from the corresponding model where prediction errors are not included (Equation 6).

[18]Note that distinguishing behavioral types is irrelevant for infrequent users who do not qualify for any discount. 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.

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 the same day of the week but from other weeks of the month.

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 is the log difference between the ridership from the indicated sample of users in that OD pair on a day in April 2015 and the ridership from the corresponding user group on the same day of the week in September 2014. For Column 1, the sample includes trips 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.

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.

[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.

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. The coefficient associated with the log average price has the "wrong" sign, which suggests that ironing has a particularly poor fit of the data.

Columns 3 and 4 report estimation results from three-price mixture models in which the final marginal and 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, are similar. 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 final 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 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.[20]

3.3.5 Discussion 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 those who have much to save from discounts. While identifying sources of such sub-optimal behavior is beyond the scope of this paper, We discuss some potential explanations here.

The first explanation is that users face substantial friction in calculating and carrying out theoretically optimal actions. Such frictions should lead to *approximately* optimal solutions, which we find no empirical support 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. However, evidence of myopia is limited.

The third explanation is gradual learning. Our results are based on the ridership in April 2015, which was only three months after the fare adjustment. Users might still be learning about the

[20] Appendix B.6 presents additional robustness checks. All evidence suggests that users are predominantly oblivious to cumulative quantity discounts.

new fare and gradually adjusting their behavior. To test this hypothesis, we replicate the same set of tests in this section using the ridership data from October 2018, which was almost four years after the fare change. We report in Appendix B.7 little evidence that users responded to the nonlinear monthly budget rationally or heuristically.

A probable 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, a small screen on the gate shows the actual charge for the trip and the remaining balance on the smartcard. In the hectic environment of the Beijing Subway, it may be challenging to back out the instantaneous discount rate, and it is even harder to predict the monthly expenditure without diligently keeping a log of subway trips.

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.

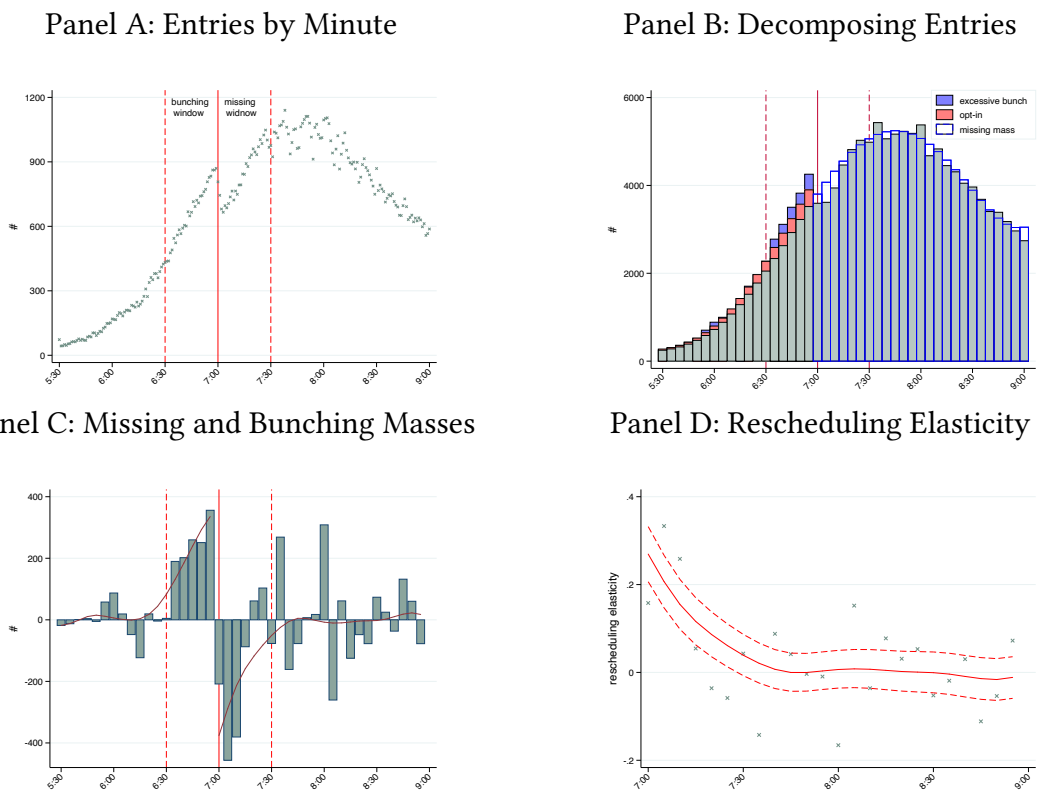
3.4 Rescheduling Elasticity

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 weekdays 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 scatter plot 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 a sharp drop in the number of entries after 7 AM. In contrast, there was no such discontinuity in September 2015, before the EBD was implemented (Appendix Figure B.13 Panel A). The discontinuity in ridership in response to the temporal price discontinuity speaks to the flexibility of users' travel schedules.

The sharp discontinuity in ridership 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 around 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^{r}_t, 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 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 is plotted with the actual ridership data and reflects 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 time cutoff, the observed number of entries with the EBD is the sum of (1) the counterfactual number of trips (represented by the green bars), (2) opt-in trips (red bars), and (3) rescheduled trips (blue bars and henceforth referred to as the *bunching mass*). To the right of the time cutoff, the observed number of entries (green bars) is the difference between (1) the counterfactual trips and (2) those rescheduled to an earlier time (hollow bars and henceforth referred to as the *missing mass*). The missing mass equals the bunching mass.

Figure 8: Early Bird Discount and 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 vertical red dashed lines indicate the rescheduling window in the baseline analysis. 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. Dashed lines represent the smoothed 95% confidence intervals (the smoothed 2.5th percentile and the smoothed 97.5th percentile from 1,000 bootstraps).

To estimate e_t^{r}, we focus on EBD stations between 5:30 AM (when most subway stations start to operate) and 9 AM on weekdays. In the baseline, we impose the demand elasticity, e^{d}, to be -0.36, which is the preferred estimate from the distance RD design (Table 2 Column 6).

We need to decide the *rescheduling windows* within which changes of travel time take place. Intuitively, the cost of rescheduling increases with the difference between the initially planned departure time and the rescheduled departure time. For the baseline, we set the *missing window* to be between 7 and 7:29 AM and the *bunching window* between 6:30 and 6:59 AM. 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. We test the robustness of the results to alternative window widths.

The estimation takes the following steps.

Step 1. Let N_t be the observed number of entries at time t and N_t^c be the counterfactual number of entries with the EBD. Entries to the left of the bunching window (between 5:30 and 6:29 AM) consist of counterfactual trips and opt-in trips. With the 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 EBD rate. Entries to the right of the missing 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 polynomial function of t. We predict counterfactual entries, $\hat{N}_t^c$, over the entire sample window (between 5:30 and 9 AM).

Step 3. Excessive bunching at time t in the bunching window is calculated as

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

$\hat{N}_t^c \cdot e^d \cdot \Delta p/p$ accounts for the opt-in trips. The number of rescheduled trips at time t in the missing window 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 ($t \geq$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 is 995, and the bunching mass is 1,263. B and M are similar, and are rather small compared with 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% of the ridership during the morning rush hours (7-9 AM).

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

Appendix B.9 includes a host of robustness checks. 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 counterfactual. Results from all those checks are remarkably similar to those in the baseline.[21]

4 Welfare Impacts

4.1 Consumer Welfare and Revenue

We evaluate the impacts of the fare structure change on revenue, consumer surplus, 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. The price for a subway trip is p_0 before the fare adjustment. The monthly budget constraint is represented by line $\bar{C}D_0$. A consumer chooses S_0 and obtains a utility level u_0.[22]

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 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 $S_{ra} < S_0$

[21] 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.

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

and $u_{ra} < u_0$.

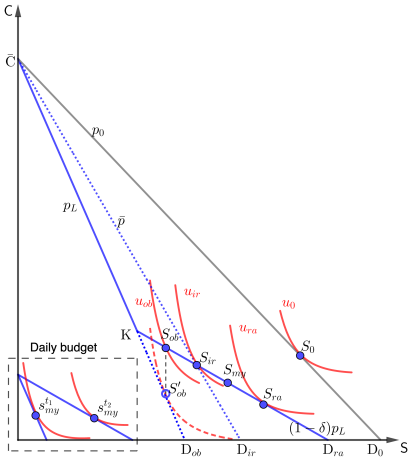
An oblivious consumer ignores the discounts and thinks, mistakenly, that the price is always p_L, and she is on the budget line $\bar{C}D_{ob}$. Under this misperception, she thinks her optimal choice is at S'_{ob}, although she actually receives discounts, and her consumption is on the budget line at 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 correctly predicts her monthly demand and is aware of the quantity discount. However, she responds to the average, $\bar{p}$, 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 polar cases of rational and oblivious, and so is her utility level u_{ir}.

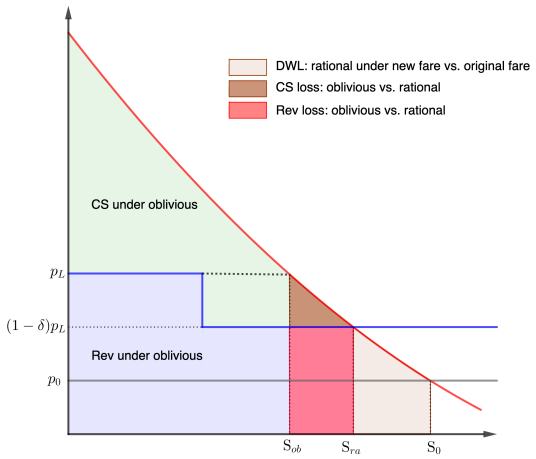
Finally, a myopic consumer responds to the current marginal price without taking the monthly budget into consideration. Her choices based on the *daily* budget are depicted in the lower-left corner of Panel A. At the start of the month, her price is p_L, and her optimal choice is $s^{t_1}_{my}$. As the month proceeds and her cumulative expenditure exceeds the cutoff, she qualifies for the discounted price of $(1-\delta)p_L$, under which her optimal daily demand is $s^{t_2}_{my}$. The monthly total of the subway trips is S_{my}, which, as we show later, is similar to S_{ir}.

Figure 9: Impacts of the Fare Structure Change on Consumer Welfare and Revenue

Panel A: Demand for Subway Trips



Panel B: Consumer Welfare and Revenue



Note: Panel A illustrates 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 with rational and oblivious consumers.

Panel B illustrates the welfare impacts of the fare structure change under two polar behavioral types: rational and oblivious. For simplicity, here we assume the *social* marginal cost of a subway trip is zero. 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.

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

Panel B also shows that it is intuitive to calculate welfare impacts. Price and discount schedules are known. With individual ridership in the pre-fare-change period, ridership under different behavioral types can be calculated using the estimated demand elasticity.

We populate the one week of the pre-fare-adjustment ridership data in September 2014 to the entire month.[23] 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). 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 across trip types within the same user type. Therefore, we assign a single demand elasticity for each user type. 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 fare structure change's impacts on the average user's ridership and welfare, assuming all users respond only to the listing price (oblivious). The average monthly number of subway trips per user declined 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 subway 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. If users respond rationally, the average user takes 16.4 trips per month under the new fare structure, 15% more than an oblivious user. Revenue would be 9% higher (63 yuan versus 57 yuan), and the deadweight loss would only be 58% lower (5.4 versus 13). The ridership

[23] We need full-month data to calculate the discount rate and how users respond to the nonlinear monthly budget. Appendix D.1 describes the simulation process that generates the pseudo data covering the entire month.

and welfare consequences of myopic or ironing users are similar (Columns 5 and 6),[24] and lie between the polar cases with oblivious or rational users.

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 user-level monthly averages.

4.2 Congestion Externality

Table 5 shows that the fare structure change led to substantial reductions in subway trips (ranging from 11% if users 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 external cost of traffic congestion (MECC) due to an additional 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

[24]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.

according to the 2014 Beijing Household Travel Survey. VOT is 23.1 yuan.[25]

T_t is the inverse of speed, measured in hours per kilometer. We use the hourly speed data from Yang et al. (2020), which are from a set of road monitoring stations on Beijing's roads in 2014.[26] $\varepsilon_t = -(\partial Speed/\partial Density) \cdot (Density_t/Speed_t)$ 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.[27]

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. We assume 50% of the reduced 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. Gu and Zou (2023) find that one passenger-kilometer by bus generates a congestion externality that is equivalent to 0.4 passenger-kilometer by car. Trips in the last category are assumed to have no impact on road congestion.

The last row of Table 5 reports the monthly external cost of traffic congestion an average subway user generates as a result of the fare structure change. If users are oblivious, the congestion externality amounts to 8.4 yuan per user per month. This is around the same magnitude as the deadweight loss (13 yuan). The negative externality will be smaller if users respond to the discounts at least heuristically. If consumers are fully rational, the 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, it only reports cases under oblivious and rational users.

Infrequent users and those who mostly travel on weekends or during weekday non-rush hours typically do not have enough trips to qualify for any discount. So alternative behavioral models make little difference in the impacts on ridership and welfare. Also, the demand elasticity for these groups is relatively small, so there are relatively small changes in ridership, substantial increases in revenue, and relatively small efficiency loss and small external cost of congestion externality.

[25] 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).

[26] Several adjustments are made to the speed data to fit our needs. Appendix D.3 describes the details.

[27] Beijing has five ring roads, which divide the city 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 each region. When a trip crosses several regions, portions of the trip are allotted to each region according to the share of the trip in each region.

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. The large decline is due to two main reasons. First, the demand elasticity for this group, at around -0.55, is the largest among all user types. Second, as frequent users, they qualify for large discounts. Ignoring incentives from discounts results in a substantial departure from optimal choices and generates large welfare losses. For the average user in this group, 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. For this group, users' behavioral responses to discounts matter significantly. If they were rational, 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 all-purpose users. The deadweight loss would have been 79% smaller if they were rational instead of oblivious. Compared with regular commuters, this group 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 reductions in subway trips cause smaller increases in road congestion.

Table 6: Distributional Impacts of the Fare Structure Change

	Infrequent users			*Weekenders*			*Weekday off-peak users*		
	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			*All-purpose users*		
	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 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

Calculations in this section show that due to inelastic demand, the fare structure change effectively raised revenue at a relatively small cost of the ridership. Consequently, the incidence is

largely shouldered by consumers. Consumer welfare loss is larger than the increase in revenue. 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 spend a lot more in percentage terms, while frequent users experience larger increases in expenditure in dollar terms.

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 users respond to the quantity discount greatly affects aggregate and distributional impacts. Being oblivious to the discounts results in fewer subway trips, which leads to lower revenue, larger consumer welfare loss, and larger cost of 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 and eliminates any concern for behavioral responses to complex pricing schedules. Second, we add to the new fare structure and expand the EBD to a peak-hour premium that applies to all stations.[28] while keeping other aspects of the new fare structure. Peak-hour premium is popular among transit systems and is often justified by the higher marginal operating costs during

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

those hours and its role in reducing system crowdedness during rush hours. The demand and rescheduling elasticities estimated in the previous sections govern user choices under alternative fare structures.

For both alternative fare structures, we calculate fare levels that would generate the same level of revenue as under the new fare structure. Revenue under the new fare structure depends on how users respond to the nonlinear monthly budget, so a fare level is calculated for each behavioral assumption. The calculation follows a double-layered iterative algorithm. 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 new fare structure. Appendix Section D.2 describes the details of the algorithm.

Table 7 presents the fare level, ridership, and welfare under the alternative fare structures. In the interest of space, the table reports cases where users are rational and oblivious. Results under myopic and ironing users are reported in Appendix Table D.1 reports

Panel A reports the revenue-equivalent fare levels. If users are oblivious, the new fare structure generates 57.3 yuan per user per month. The alternative flat rate needs to be set at 3.78 yuan 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 and 2.26 yuan during off-peak hours. To make fare levels comparable across alternative pricing schedules, the last row in Panel A reports the listing price per kilometer under the trip composition in September 2014. 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; Panel C reports associated welfare changes relative to the original 2-yuan flat rate. All numbers reported in these two panels are in per user per month terms.

The average oblivious user takes 14.3 trips and spends 57.3 yuan under the new fare structure. Compared with the 2-yuan flat rate, while the revenue increases by 20.5 yuan, consumer welfare declines by 33.5 yuan, resulting in a deadweight loss of 13 yuan. It also generates a cost of congestion externality equaling 8.4 yuan.

In this case, a revenue-equivalent 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

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 users are oblivious or rational to the quantity discounts. Pre-fare-adjustment monthly ridership patterns are based on data from September 2014. The Deadweight loss is users' 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.

congestion is 29% smaller. The flat rate outperforms the new fare structure because oblivion to discounts disproportionately hurts frequent users, who derive high values from subway trips and are more likely to travel during peak hours, and because diverted peak-hour subway trips cause larger negative externalities on surface roads.

However, if users are rational, the new fare structure outperforms the flat rate. Compared with the new fare structure, the average rational user takes 11% fewer trips under the flat rate. Consumer welfare loss, deadweight loss, and the cost of congestion externality are 21%, 122%, and 52% higher, respectively, under the flat rate. The new 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. However, it performs worse in practice as users do not rationally respond to the incentives.[29]

Adding a peak premium to the nrew fare structure results in larger welfare losses and larger congestion externalities regardless of user's behavioral types (comparing Columns 2 and 4, and Columns 5 and 7). This is because peak-hour pricing further penalizes frequent users, who have high-valued trips during rush hours. Appendix Figure D.2 shows that peak-hour pricing is the most effective in reducing rush-hour system load. Still, due to the low rescheduling elasticity,

[29] Appendix Table D.2 reports the distributional impacts of alternative fare structures on different user types.

the peak-hour pricing does not achieve this by moving trips to less busy times. Rather, it likely diverts most of the trips to surface roads, which results in high costs of congestion externality.

6 Conclusion

The design of transit fare structure 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 a 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.

Users do not seem to respond to the cumulative quantity discounts. We can soundly reject that users are forward-looking and optimize on the monthly budget. Evidence also does not support users 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 for a better design of fare structures. The new 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 associated with a complex fare structure, would be preferred if users 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 could generate large negative externalities by worsening surface road congestion.

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," *Mimeo*.

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," *Working paper*.

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

Cahana, M., N. Fabra, M. Reguant, and J. Wang (2022): "The distributional impacts of real-time pricing," CEPR Discussion Paper No. DP17200.

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, M. K., P. E. Rossi, J. A. Chevalier, and E. Oehlsen (2019): "The value of flexible work: Evidence from Uber drivers," *Journal of Political Economy*, 127, 2735–2794.

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." Unpublished.

Hahn, R. W., R. D. Metcalfe, and E. Tam (2023): "Welfare Estimates of Shifting Peak Travel," Tech. rep., National Bureau of Economic Research.

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.

Holmgren, J. (2007): "Meta-analysis of public transport demand," *Transportation Research Part A: Policy and Practice*, 41, 1021–1035.

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*.

Larcom, S., F. Rauch, and T. Willems (2017): "The benefits of forced experimentation: Striking evidence from the London underground network," *The Quarterly Journal of Economics*, 132, 2019–2055.

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.

Litman, T. (2004): "Transit price elasticities and cross-elasticities," *Journal of Public Transportation*, 7, 37–58.

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," *Mimeo*.

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.

Yang, N. and Y. Long Lim (2018): "Temporary incentives change daily routines: Evidence from a field experiment on Singapore's subways," *Management Science*, 64, 3365–3379.

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

Fare Structure and the Demand for Public Transit Online Appendix (Not for Publication)

Table of Contents

A Data Summary **2**

B Additional Robustness Checks **3**

B.1 Robustness Checks for RD Estimations of the Demand Elasticity 3

B.2 Estimating the Bunching Elasticity 6

B.3 Demand Elasticity from User-level Estimations 10

B.4 Tests of Rationality with Regular Commuters 13

B.5 Additional Tests of Myopia 13

B.6 Additional Results on the Composition of Behavioral Types in Responses to Quantity Discounts 16

B.7 Evidence of Learning 17

B.8 Information Display at the Gate 24

B.9 Robustness Estimations of the Rescheduling Elasticity 25

C Details of User Classification **32**

C.1 Additional Details of the K-means Clustering Algorithm 32

C.2 Ridership by Time and User Type 33

C.3 Classification of Users in September 2014 34

D Details of Welfare Calculations under the Current and Alternative Fare Structures **35**

D.1 Populating Data in the week of September 15, 2014, to Full Month 35

D.2 Algorithms to Calculate Ridership and Counterfactual Fares 36

D.3 Extrapolating Road Speed 37

D.4 Additional Results on Ridership and Welfare under Alternative Fare Structures . 39

E Heterogeneity and Welfare Incidences by User Skills **45**

E.1 Data Description and Validation 45

E.2 Demand Elasticity by Skill 50

E.3 Welfare Impacts by Skill 51

A Data Summary

The trip-level smartcard data are provided to us by the Beijing Institute of City Planning (BICP). The BICP obtains samples of ridership data from the subway company to facilitate their work in city planning. Typically, the BICP requests the universe of the trip-level data captured by smartcards in one week in every season. We have access to 13 data waves ranging between September 2014 and October 2018. Two data waves, one in April 2015 and the other in October 2018, contain the universe of trips in the entire month. The universe of ridership data in the entire month allows us to analyze how consumers respond to the non-linear monthly budget. The week of September 15, 2014, is the only wave before the fare structure change.

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. The three waves of data include two full weeks and one full month. The baseline analysis focuses on changes in ridership between the week of September 15, 2014, and the month of April 2015. Ridership data in the week of September 12, 2016, are used to estimate the rescheduling elasticity from the early-bird discounts, which were implemented in 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] 4/6/2015 (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 the sample period, during which the ridership pattern could be different. When a national holiday occurs, workdays and weekends are reorganized to make non-workdays in the official holiday contiguous. Such reorganizations 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 included three days, which included 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, individual firms could decide whether to extend the holiday to include that Sunday, and many did. In our analyses, we flag all those affected days as potentially affected by the holiday. In the case of the 2016 Mid-autumn holiday, we treat all days between September 15 (Thursday) and September 18 (Sunday) as non-work 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.[1] 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 with waiting time inclusive. 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.11.

B Additional Robustness Checks

B.1 Robustness Checks for RD Estimations of the Demand Elasticity

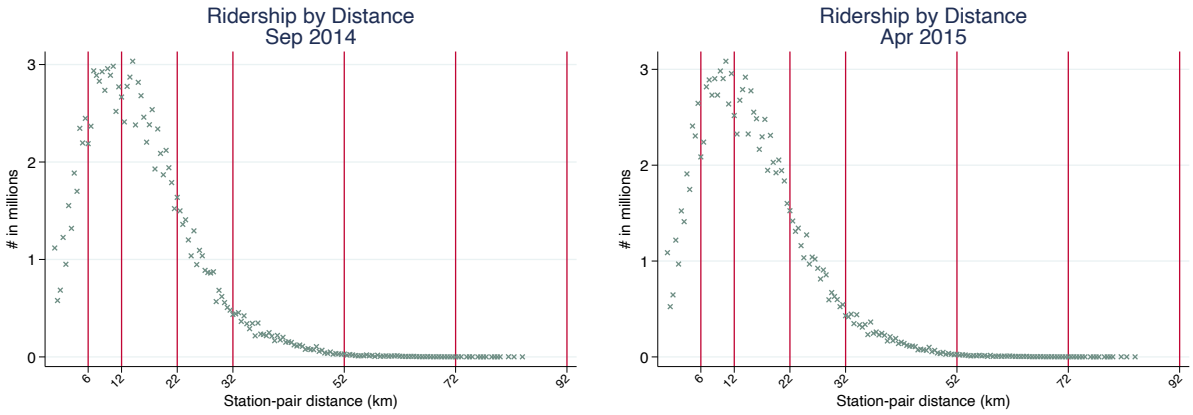
Ridership by station-pair distance bins. Figure 1 shows the log *changes* in ridership between September 2014 and April 2015 in station pairs by 500-meter distance bins. There are visually sharp discontinuities around distance cutoffs. Appendix Figure B.1 plots ridership *levels* by station pairs in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

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.

[1] The official statistics report an average daily ridership of around 9 million. The official statistics define a ride as being at the trip-line level. For example, if a subway trip involves two transfers between three lines, the entire trip is regarded as three rides in the official statistics. On average, each subway trip involves 2.4 lines. Our data cover the near universe of all subway trips.

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.

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



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

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 polynomial function $f(.)$. This approach 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 non-parametric 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.[2] Appendix Table B.1 reports the results. The demand elasticity estimates are remarkably similar to those from global polynomials, reported in Table 2.

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 (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. So there are in total 128 regressions (four

[2]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 that equals one if the station pair is located on the right side of the cutoff.

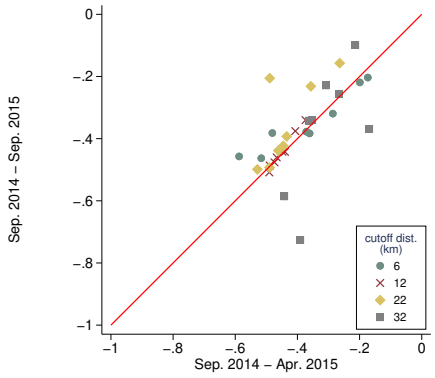
Table B.1: Demand Elasticity Using Local Linear Regression Discontinuity Design

	(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]	-

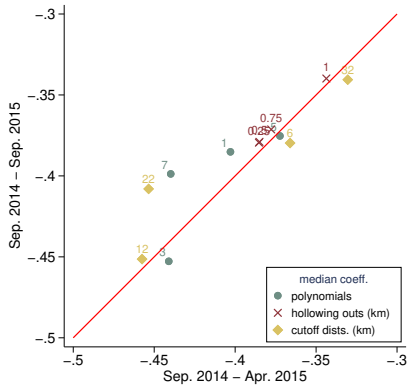
Note: The table reports results from estimating Equation 1 using regression discontinuity models with local linear regression, triangular kernel, and optimal bandwidth. Robust standard errors are in parentheses. The dependent variable is the log change in ridership in the same OD pair between September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

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 16 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.

degrees of polynomials, four widths of the hollowing out regions, four distance cutoffs, and two time ranges).

Appendix Figure B.2 summarizes the results from those robustness checks. Panel A reports the results of all 128 estimations. Most estimates are tightly distributed except for a few estimations from the 32-km cutoff. Estimates around the 32-km cutoff are less precise, 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 estimate. 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 $p(1+\gamma)$ while that of OD_L remains unchanged. So the change in the difference in price $\Delta p/p = \gamma$. 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 that are bunched to OD_L in exchange for a one-percent saving on fare. The example in Table B.2 shows that $e = \frac{-x}{Q\gamma}$, $e^b = -\frac{y}{Q\gamma}$, and $e^{RD} = \frac{-x-y}{Q\gamma} - \frac{y}{Q\gamma} = \frac{-x-2y}{Q\gamma}$. 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	$p(1+\gamma)$
# 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 $p(1+\gamma)$. 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 September 2014 and April 2015, we present a way to directly estimate e^b and thus recover the demand elasticity e as $e = e^{RD} - 2e^b$.[3] At each cutoff, we estimate the following equation:

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

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

[3] 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, the estimates are likely the upper-bound of e^b. The same methodology applies to all subway users, and the results are similar.

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 w.

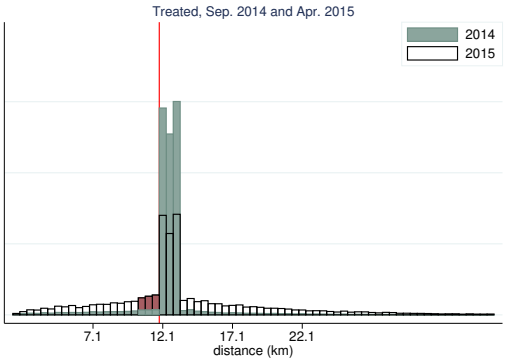
The control group is defined as those who have the majority of their trips in 2014 within a pre-specified "control window" $[c + d, c + d + w]$. OD pairs in the control window are at least d km longer than the distance cutoff but shorter than the next distance cutoff. The control group faces the same price as the treated group but is assumed to have no incentive to bunch because doing so would require them to shorten their trips substantially. 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^o_{i14} is the number of trips from a treated (control) rider i in 2014 that falls in the *treated* (*control*) window, and ΔN^b_i is the change in the number of trips in the corresponding *bunching* window between 2014 and 2015.[4] $\Delta N^b_i / N^o_{i14}$ 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 as the control group. 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.

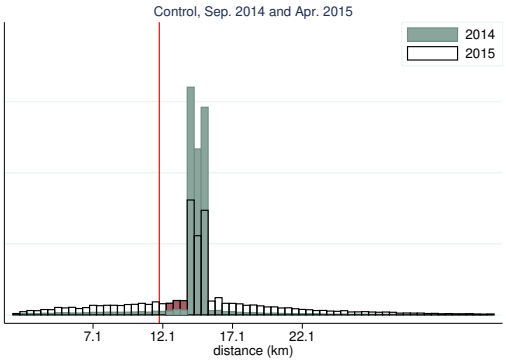
[4]The number of trips in the week in September 2014 is converted to the number of a full month by multiplying 30/7.

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. The treated group includes those who had the majority of trips in distance bins right above the cutoff in 2014, while the control group includes those who had the majority of trips in distance bins a little farther above the cutoff in 2014. Blue bars indicate the ridership in each distance bin in 2014, and the white bars indicate the ridership in 2015.

Figure B.3 graphically illustrates the intuition of the strategy to estimate e^b. Focusing on the 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. The flattening out of the distribution reflects natural regression towards the mean. It highlights the need for a control group to isolate 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.

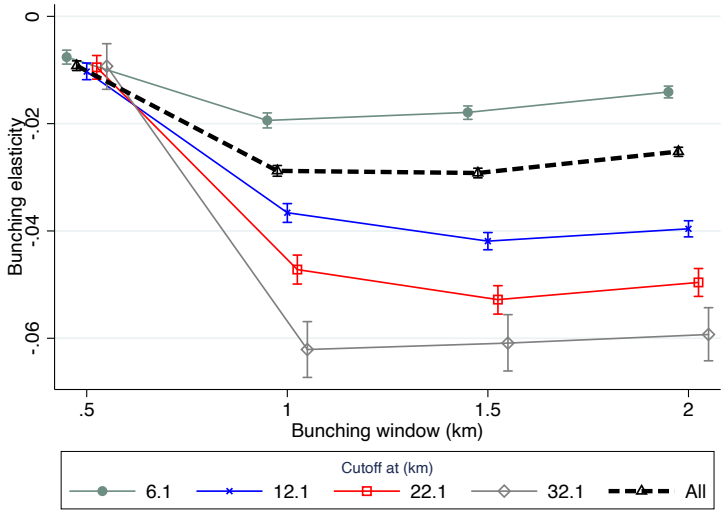
Table B.3: Bunching Elasticity

	(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 treated window or the control window). The window width is 1.5 km. Robust standard errors are in parentheses.

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 Windows



Note: Estimations of Equations B.1 and B.2 with varying bunching window width w at each distance cutoff as well as jointly with all cutoffs. Vertical bars indicate 95% confidence intervals associated with each estimate.

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

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

The outcome variable is the same as that in Equation B.1. T_{ic} indicates treatment status of user 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 (Column 5 of Table B.3), which is similar in magnitude to the weighted average of bunching elasticity estimated separately for each cutoff (Columns 1 through 4). Again, the result is not sensitive to w as long as it is above 1 km (Figure B.4). Take $e^{RD} = -0.387$ (Table 2 Column 5) and $e^b = -0.029$, we recover the bunching-adjusted 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 an alternative 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 data and identification challenges. Nevertheless, the estimate from this approach is likely to be a lower bound of the true elasticity. We show that the estimated elasticity is similar, though as expected, somewhat lower than our preferred specification.

We estimate the following equation using the 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 21^{st}. $\Delta \ln D_i$ is the log 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 the weekly ridership in September 2014 into monthly by multiplying 30/7. We

add one to both numbers of trips and the total distance to avoid the problem of taking a logarithm over zero. The results are similar if inverse hyperbolic sine is used instead of the logarithm, or levels are used as the outcome variable and then the coefficient is converted 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 had taken in September 2014 after the fare adjustment. We measure the change in expenditure based on the monthly 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, users mostly react to the listing price and do not respond to discounts in any way.

e^{card} can be interpreted as the demand elasticity. To be consistent with the baseline estimate of the demand elasticity that represents the aggregate ridership, each user is weighted 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. Both sets of polynomials are always included in all regressions. 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 will be only marginally 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 of the user 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 change in the fare structure. Indeed, idiosyncratic demand shocks can be substantial. 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. Therefore, the OLS estimate of Equation B.3 results in a lower bound of the demand elasticity.

The conventional approach to address the classical measurement error 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 of the monthly demand 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, not reported here, 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.

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 the user in the initial period. In Column 2, each observation is weighted by the user's total distance from all trips during the initial period. Robust standard errors are reported in parentheses.

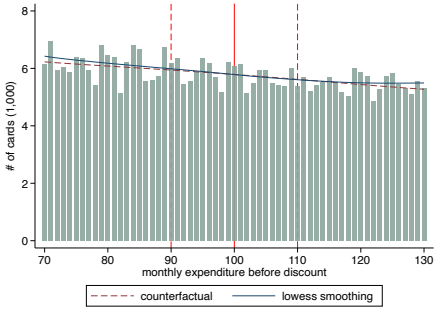
Column 1 of Table B.4 reports the demand elasticity for the number of trips is -0.27. It is slightly smaller than the baseline result from the OD-pair RD regressions, which is consistent with the attenuation bias due to classical measurement error. This number can be seen as the lower bound of the demand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

distances, the impact on passenger-kilometers is larger than that on the number of trips.

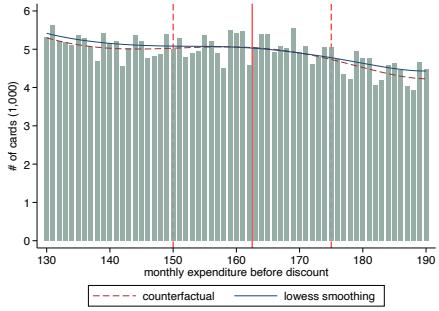
B.4 Tests of Rationality with Regular Commuters

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 around the first two concave kinks at 100 and 162.5 yuan, respectively. In both graphs, the solid vertical line indicates the kink point and 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 are 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.

In Section 3.3.2, we show there is no evidence that consumers rationally or heuristically 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.

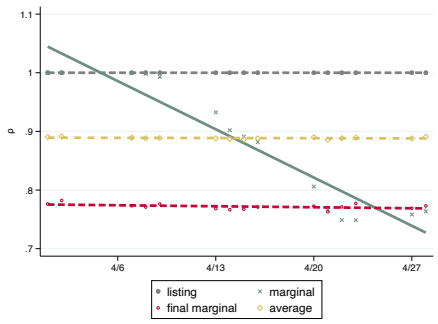
One may suspect that frequent users with a regular travel pattern are more likely to be able to predict their monthly demand and rationally respond to it. Figure B.5 repeats the same density analysis among those we classify as regular commuters. Again, the empirical monthly budget has no visual or statistical evidence of hollows or dents around the first two nonconvex kinks. Inspecting the density in the neighborhood around the third, context, kink point (at the pre-discount monthly expenditure of 662.5 yuan) results in the same conclusion. We do not present the graph here because few users are in the neighborhood.

B.5 Additional Tests of Myopia

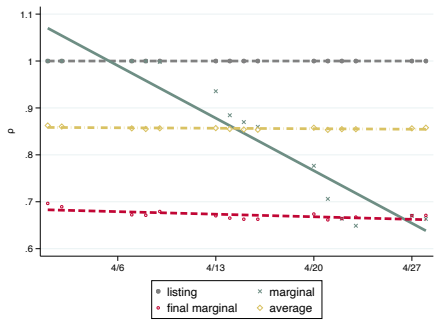
Section 3.3.3 shows limited evidence for myopia. Here, we present some additional checks.

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

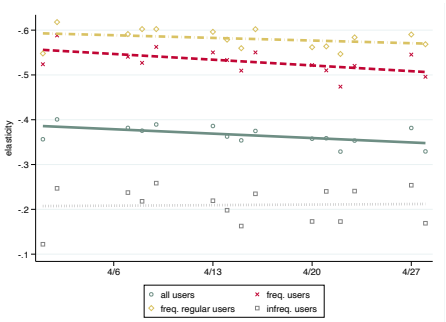
Panel A: RD in Prices, Frequent Users



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-discount 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 the date of April 2015, assuming users respond only to the listing price. The regression equation is described in Equation 4. Each of the four series corresponds to a different sample, while the linear line in the corresponding color is the linear fit of those estimates. The four samples are (1) trips from all users (The fitted line has a slope of 0.0014 and a robust standard error of 0.0007), (2) trips from frequent users (slope=0.0018 and s.e.=0.0009), (3) trips from frequent users who have regular commuting patterns (slope=0.0008 and s.e.=0.0009), and (4) trips from less frequent users (slope=-0.00018 and s.e.=0.0017). Lines in the corresponding color are linear fits of those coefficients. 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 for clean illustration.

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-discount 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 our estimates. 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 0.002. This is evidence that the underlying demand elasticity, as a structural parameter, does not change over time.[5]

We then focus on trips from frequent users with a pre-discount 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. While they may take trips that land at different sides of the distance cutoffs, there will be no 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 by users' travel patterns). One may argue that users with regular travel patterns 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 final marginal price, and monthly average price all lie in flat lines for both groups, discontinuities in the instantaneous marginal price are in downward-trending lines. Panel C reports $\hat{e}_t^{\text{listing}}$ from estimating Equation 4 for both groups. Red crosses represent estimates from frequent users, while yellow diamonds represent those from regular commuters. 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.

[5]Note that infrequent users have a smaller demand elasticity, which is consistent with Figure 3. In Panel C, the $\hat{e}_t^{\text{listing}}$'s from infrequent users lie below those from frequent users.

B.6 Additional Results on the Composition of Behavioral Types in Responses to Quantity Discounts

Appendix Table B.5 reports results from estimating versions of the statistical mixture model (Equations 5 and 7). The specifications correspond to those in Table 4, but the models are estimated using the Original Least Squares (OLS). The results are quantitatively similar to those in Table B.5, which are estimated using the Two-stage Least Squares (TSLS) 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 estimator. Each observation is an OD pair by date. 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 on the same day of the week in September 2014. In Column 1, the sample includes ridership from all users. In Columns 2 through 6, the sample includes trips from frequent users with a pre-discount monthly expenditure of more than 70 yuan. In Columns 7 and 8, the sample includes trips from frequent users who have 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 that estimated in the corresponding specification for which optimization errors are not accounted. All regressions include a 5^{th}-order polynomial of OD-pair distance. 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 raced 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

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 one alternative behavioral assumption. Each observation is an OD pair by date. The dependent variable is the log difference between the ridership from the specific sample of users in that OD pair on a day in April 2015 and the ridership from the corresponding user group on 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. Models in Panel B are estimated using the TSLS, where the 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.

frequent users, and (3) trips from frequent users who also have a regular travel pattern. There is overwhelming evidence for all three samples that users respond only to the listing price. In all estimations, the implied demand elasticity remains consistent for the corresponding sample of users. Both OLS and IV estimations yield quantitatively similar results.

B.7 Evidence of Learning

Sample and user classification. The main analysis uses ridership data from April 2015. We find no evidence that users respond in any way to the nonlinear monthly budget. However, April 2015 is merely three months after the fare structure change. Users might still be in the process

of familiarizing themselves with the new fare structure and formulating their optimal responses. The learning process could take some time.

We test the relevance of the learning hypothesis by using the full-month ridership data from October 2018, which is almost four years after the fare structure change took place. We replicate the set of empirical tests of behavioral responses to the nonlinear budget constraint, as presented in Section 3.3. Overall, we find remarkably similar results between April 2015 and October 2018. Therefore, learning cannot explain the lack of any response to the nonlinear monthly budget.

Frequent and regular users have stronger incentives to learn and arguably lower costs to adjust. To explore potential heterogeneity, users are first classified by their travel patterns. We use the same set of variables described in detail in Appendix C as predictors. There was a one-week holiday (the National Day holiday) at the beginning of the month, which substantially reduced the number of workdays in that month. Holidays are treated as weekends. Despite that, the composition of cards is remarkably similar to those of users in April 2015. Table B.7 reports the composition and characteristics of the classified categories. About half of all users are infrequent subway users with, on average, less than two trips each month. Weekenders and weekday off-peak users each account for 16-17% of all users, and an average user in either category has a monthly ridership of about eight trips. The remaining 20% of cards belong to frequent users who ride the subway for commutes. 1.15 million such "rush-hour commuters" mostly use the subway for commuting during peak hours, while another 1.36 million "all-purpose users" also use the subway during other times and for other purposes.

Figure B.7 shows the makeup of the daily ridership by the five types of users, separately for weekdays and weekends/holidays. Similar to the patterns in April 2015 (Figure C.1), ridership during weekdays has two peaks during morning and afternoon rush hours. Rush-hour commuters and all-purpose users account for the bulk of those trips. The numbers of trips—for example, around 600,000 during morning rush hours and 500,000 during afternoon rush hours—are also similar between the two periods.

Density tests for rationality. The cumulative quantity discounts create non-convex kinks in the monthly budget. Section 3.3.2 shows how such kinks may lead to non-smoothness in the user distribution in terms of monthly ridership. Specifically, if users are rational, we expect to see a dent in the density of user distribution around the kink points. The depth and sharpness of the dent depend on the distribution of the underlying demand elasticity and the magnitude of optimization frictions users face. In the ideal case in which all users have the same demand elasticity and no optimization frictions, the density of users near the kink point is zero. A wider hole corresponds to a larger demand elasticity.

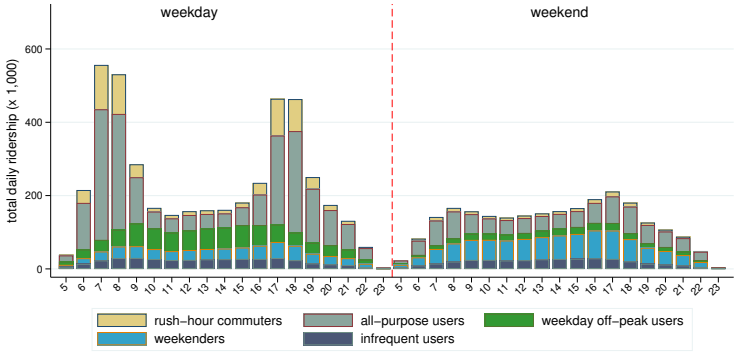
Figure B.8 plots the empirical density of user distribution around the two non-convex kinks: the first is at 100 yuan in pre-discount monthly expenditure, while the second is at 162.5 yuan.

Table B.7: Card Classification and Characteristics, October 2018

	infreq. users	weekenders	weekday off-peak users	rush-hour commuters	all-purpose users
	(1)	(2)	(3)	(4)	(5)
# of users (million)	6.07	2.15	2.00	1.15	1.36
share of users	0.48	0.17	0.16	0.09	0.11
# of rides (monthly)	1.81	8.23	8.89	12.74	38.61
	(0.74)	(5.00)	(5.98)	(7.82)	(10.20)
total distance (km)	30.71	134.95	141.71	186.26	638.04
	(22.75)	(105.06)	(121.43)	(158.43)	(343.22)
share of rides during					
weekday AM rush	0.15	0.09	0.10	0.40	0.35
	(0.29)	(0.12)	(0.14)	(0.21)	(0.15)
weekday PM rush	0.14	0.13	0.11	0.33	0.26
	(0.29)	(0.14)	(0.14)	(0.21)	(0.15)
weekday non-rush	0.33	0.15	0.61	0.18	0.25
	(0.42)	(0.21)	(0.18)	(0.15)	(0.19)
weekend	0.38	0.63	0.18	0.08	0.15
	(0.47)	(0.21)	(0.18)	(0.16)	(0.15)
# of weekdays traveled	1.01	2.82	4.47	7.45	17.04
	(0.69)	(2.20)	(2.85)	(4.37)	(2.68)
location bin HHI	0.83	0.36	0.38	0.64	0.62
	(0.25)	(0.21)	(0.22)	(0.26)	(0.25)
OD location bin concen. rate	1.36	2.10	2.26	4.95	10.58
	(0.52)	(1.33)	(1.52)	(3.40)	(8.74)

Note: Users are classified into five groups based on their travel patterns in the month of October 2018 using a $K-$means clustering algorithm. The table reports the summary statistics of travel patterns for each user category. Standard deviations are in parentheses. See Appendix C for details of the clustering algorithm.

Figure B.7: Trips Composition by Time and User Type, October 2018

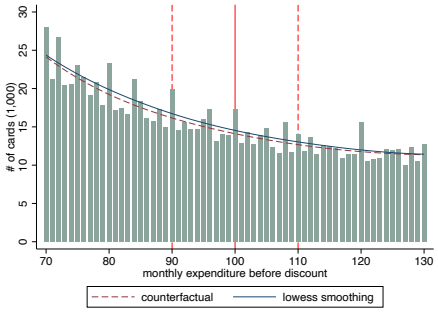


Note: Users are classified into five categories based on their travel patterns in October 2018 using a K-means clustering algorithm. The graph shows the composition of trips by card type and by day and time.

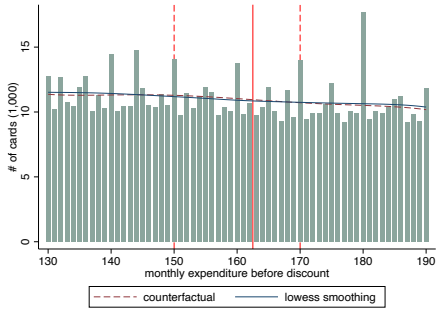
Figure B.8: Density Distribution around Non-convex Kinks, October 2018

Panel A: Full Sample

First Kink at 100 *yuan*

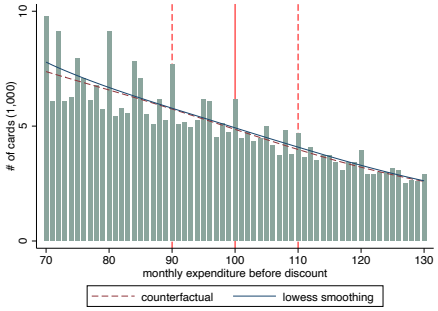


Second Kink at 162.5 *yuan*

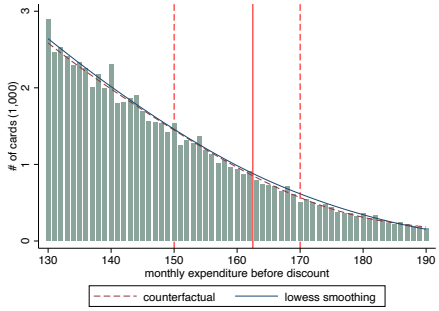


Panel B: Regular Commuters

First Kink at 100 *yuan*



Second Kink at 162.5 *yuan*



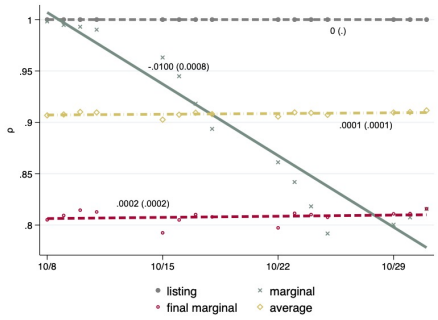
Note: Data are from the full-month ridership records in October 2018. The discount schedule creates three kinks in the budget line. Here we show the distribution of pre-discount expenditure at the first two kinks at 100 and 162.5 yuan, respectively. Graphs in Panel A include for all users. Graphs in Panel B include frequent users who have regular travel patterns. In all graphs, the solid vertical line indicates the kink point, and 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.

For each kink point, we plot user density in a 60-dollar neighborhood. Panel A plots the density distributions for all users. Notice that infrequent users are unlikely to have a monthly expenditure high enough to appear in the neighborhood of either kinked point. So, those included in the analysis are necessarily frequent subway users. Frequent and regular commuters have a stronger incentive to learn to respond to the non-linear budget and arguably have a lower cost to do so due to their predictable commuting patterns. Panel B plots the density distribution among those regular commuters. For both groups, however, there is no visual or statistical evidence of even a shallow dent around either kink point. After almost four years of the fare structural change,

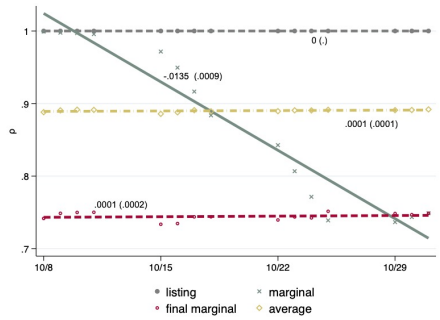
we can statistically reject the hypothesis that a meaningful mass of subway users respond to the kinked budget constraint quasi-rationally.

Figure B.9: Tests for Myopia by Sub-samples, Ridership in October 2018

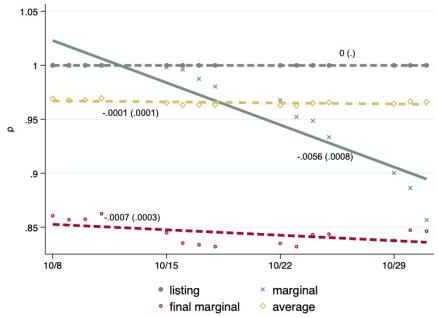
Panel A: RD in Prices, All Users



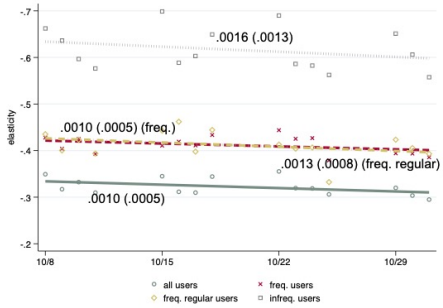
Panel B: RD in Prices, Frequent Users



Panel C: RD in Prices, Frequent and Regular Users



Panel D: Elasticity Implied by Listing Price



Note: Panels A through C plot regression discontinuity estimates (ρ_t) of perceived prices under various behavioral assumptions by each workday between Monday and Thursday in the month of October 2018. The regression equation is described in Equation 3. In Panel A, the sample includes trips from all users. In Panel B, the sample includes frequent users with pre-discount monthly subway expenditures of more than 70 yuan. The sample in Panel C includes trips from frequent users with regular commuting patterns. The fitted linear lines in each graph indicate time trends of ρ_t under various behavioral assumptions. The slope and the associated standard errors are marked on the graph. Panel D plots regression discontinuity estimates of the demand elasticity for each date in the month of October 2018, assuming users respond only to the listing price. The regression model 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 regular commuting patterns, and (4) trips from less frequent users. Lines in the corresponding color are linear fits of those coefficients. The slope and the associated standard errors are marked on the graph. In all regressions, 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 for clean illustration.

Table B.8: Mixture Model and the Composition of Behavioral Types, Ridership in October 2018

Polynomials of log final marginal p				X	
Polynomials of log avg. p					X
Panel A: All users	(1)	(2)	(3)	(4)	(5)
Log listing p	-0.250	-0.230	-0.204	-0.224	-0.216
	(0.079)	(0.078)	(0.078)	(0.084)	(0.084)
Log instan. marginal p	-0.001	-0.000	0.002	0.015	0.015
	(0.007)	(0.007)	(0.007)	(0.004)	(0.004)
Log final marginal p	0.060	0.061		-	
	(0.021)	(0.021)		-	
Log avg. p	0.024		0.027		-
	(0.007)		(0.008)		-
e constrained at				-0.168	-0.176
Panel B: Freq. users	(1)	(2)	(3)	(4)	(5)
Log listing p	-0.114	-0.142	-0.190	-0.151	-0.193
	(0.089)	(0.088)	(0.087)	(0.071)	(0.087)
Log instan. marginal p	-0.023	-0.024	-0.027	-0.027	-0.031
	(0.006)	(0.006)	(0.007)	(0.004)	(0.007)
Log final marginal p	-0.103	-0.105		-	
	(0.017)	(0.017)		-	
Log avg. p	-0.032		-0.036		-
	(0.005)		(0.006)		-
e constrained at				-0.271	-0.253
Panel C: Freq. and regular commuters	(1)	(2)	(3)	(4)	(5)
Log listing p	-0.220	-0.212	-0.211	-0.241	-0.230
	(0.090)	(0.089)	(0.089)	(0.085)	(0.085)
Log instan. marginal p	0.002	0.002	0.002	0.030	0.030
	(0.004)	(0.004)	(0.004)	(0.007)	(0.007)
Log final marginal p	0.011	0.011		-	
	(0.006)	(0.006)		-	
Log avg. p	0.009		0.009		-
	(0.004)		(0.004)		-
e constrained at				-0.200	-0.201

Note: The table reports results from estimating various versions of the mixture model. Each observation is an OD pair by date. The dependent variable is the log difference between the ridership from the specific sample of users in that OD pair on a day of October 2018 and that from the corresponding user group 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. The number of OD-pair-by-day observations is 1.56 million in Panel A, 1.35 million in Panel B, and 1.20 million in Panel C.

Tests of myopia. Using the method described in Section 3.3.3 and Appendix B.5, we test whether there is evidence for users responding to the instantaneous marginal price. The intuition of the test is briefly recapped as follows. A unique feature of the contemporaneous marginal price is that it changes over time, while the other three are constant. As a user accumulates expenditure throughout the course of the month, she gradually qualifies for a larger discount before maxing out. Compared with users who consistently take trips just below the distance cutoff, those who consistently take trips just above the cutoff qualify for discounts earlier and quality for higher discounts. This generates a *decreasing* discontinuity over time in the instantaneous marginal price across the distance cutoff. We assume that the demand elasticity is a constant. If in reality, users respond to the instantaneous price, estimating the demand elasticity using the discontinuity in the listing price would over time *over*-state the actual perceived price difference, leading to an increasingly *under*-estimated demand elasticity.

To the extent that different types of users may have different demand elasticities, the test is carried out separately for four user groups: (1) all users, (2) frequent users with a monthly expenditure of more than 70 yuan, (3) frequent users with a regular commute pattern, and (4) infrequent users. The last group serves as a placebo because they never come close to qualifying for any discount, and behavioral responses are irrelevant to them. Panels A through C of Figure B.9 plot discontinuities in various prices relative to that in the listing price (ρ_t's in Equation 3), separately for different user groups. As in the analysis using the April 2015 data, we exclude Fridays, weekends, and holidays. Because the first week of October 2018 was a long holiday, the plots start on October 8th, and have 15 observations in each series. The plot for infrequent users is not shown because the coefficients are trivial for this group – all coefficients are equal to one because by definition, no one in this group qualifies for any discount. We fit a linear line for each series; the slope of each fitted line and the associated standard error are reported in the graph.

For all three groups, the discontinuities in the listing price, month-end final marginal price, and the monthly average price are all flat over the course of the month. This is expected and confirms that in the 15 workdays of the month, user composition in a given origin-destination pair remains largely stable. Discontinuities in the instantaneous marginal price are on a downward-sloping curve. At the beginning of the month, no one qualifies for any discount; the discontinuity in the marginal listing price is the same as that in the listing price (so the ρ_t is equal to 1). By definition, by the end of the month, the instantaneous marginal price is the same as the month-end marginal price. The estimates of ρ_t's from the two series confirm this observation.

Panel D plots the estimated demand elasticities using the listing price (Equation 4), separately for each user group. Linear fits are plotted for each series with the slope and the associated standard error marked on the graph. Notice that the vertical axis labels are in reverse order, so a positive slope is associated with a downward-sloping line. All linear fitted lines have slopes that

are statistically indifferent from zero. The demand elasticity for the whole sample, estimated using the listing price, is -0.33 on the first workday of the month and is -0.31 by the end of the month, a 6% decline in magnitude. In contrast, the discontinuity in the contemporaneous marginal price decreases by about 25% (Panel A). If users respond to the contemporaneous discount rate, using the listing price would be correct to recover the demand elasticity at the beginning of the month but would underestimate the true demand elasticity by about 25% by the end of the month. We would expect a much steeper slope of 0.0036 (0.33*0.25/(31-8)). The estimated slope (0.001 with a standard error of 0.0005) can reject a slope of 0.0036 at the 1% significance level. Therefore, myopic users are unlikely to be a substantial group among subway riders in October 2018. Using different samples with trips from frequent users and regular commuters leads to the same conclusion.

Estimating the composition of behavioral types. Finally, we directly estimate the composition of user behavioral types with a statistical mixture model (Equation 5). Following the approach introduced in Section 3.3.4, we estimate various versions of the model with three different samples: all users, frequent users, and frequent users with regular commuting patterns. To break the mechanical correlation between the same-day demand shock and qualified discount rate, we instrument the instantaneous marginal price, the month-end final marginal price, and the monthly average price by predicted counterparts where the same-day ridership is replaced by the same user's average ridership on the same day of the week but in other weeks of the month. We include both workdays and non-workdays (weekends and holidays) in the regression.

Table B.8 Panel A reports results from estimating the mixture model with the full sample of users. Columns 1 through 3 show that the effect is loaded on the log listing price in four-way and three-way mixture models, suggesting that subway riders almost entirely consist of those oblivious to the monthly discounts. To allow for imperfect optimization, we include polynomials of the log month-end final marginal price and the log monthly average price in Columns 4 and 5, respectively ((Equation 7)). Because we lose a degree of freedom in those specifications, the demand elasticity is constrained at the level estimated from the corresponding model without polynomials. The results are similar to the previous columns.

Panels B and C report results with frequent users and regular commuters. For both groups, we still find that oblivious users make up the vast majority. There is no evidence of any substantive fractions of users being myopic, ironing, or rational.

B.8 Information Display at the Gate

Figure B.10 illustrates the information available to subway riders when they exit a station through a turnstile. As a passenger taps her smartcard (the photo shows a more recent version in which

a virtual smartcard is installed on a smartphone) on the reader installed on the turnstile, a small screen next to the reader displays the "fare deducted," which is the actual charge for the trip after applying the discount, If the passenger knows the listing price of the trip, she can back out the *instantaneous* discount rate she receives for the trip. The screen also displays the "remaining balance" on the card. No other information is displayed.

Figure B.10: Information Display at the Gate

B.9 Robustness Estimations of the Rescheduling Elasticity

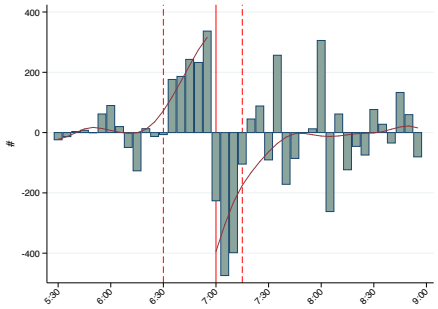
Varying the width of the rescheduling window. This section provides robustness checks to the estimation of the rescheduling elasticity introduced in Section 3.4. The first set of robustness checks involves varying the width of the rescheduling window. The choice of the rescheduling window width faces a tradeoff. 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. Therefore, we would like to set the window just wide enough to capture all rescheduled trips. Whether a window is sufficiently wide is verifiable by inspecting whether all rescheduled trips are wrapped within the window.

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

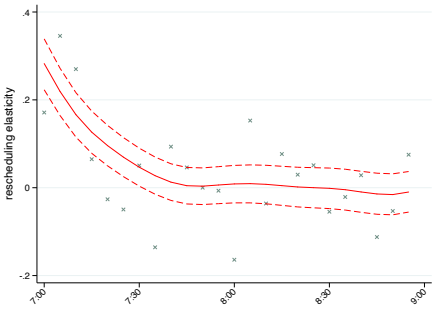
Figure B.11: Robustness Estimations of the Rescheduling Elasticity - Varying Bunching and Missing Windows

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

Missing Mass and Bunching

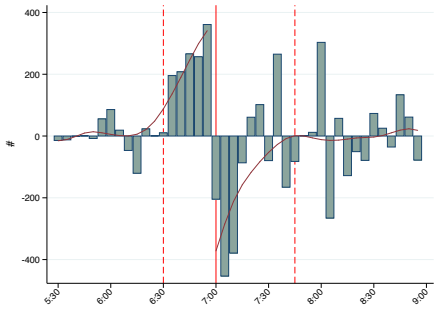


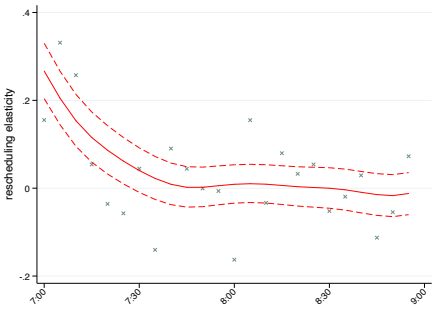
Rescheduling Elasticity



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

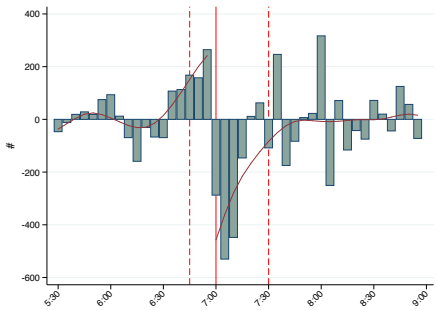
Panel B: Rescheduling Window between 6:30 AM and 7:45 AM

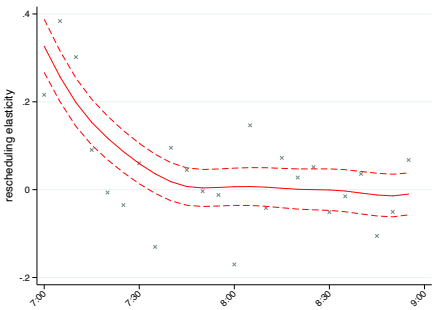




B=1,298, *M*=1,028, Counterfactual number of trips in the missing window=42,924

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

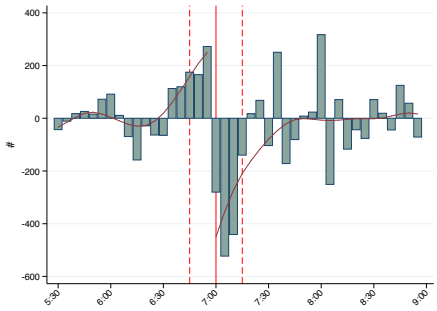




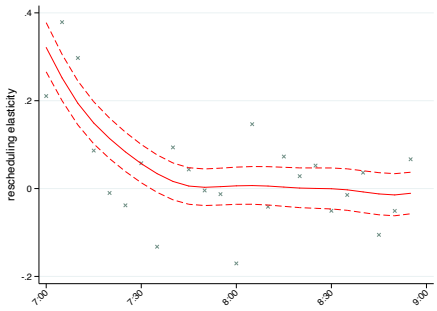
B=590, *M*=1,372, Counterfactual number of trips in the missing window=27,807

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

Missing Mass and Bunching

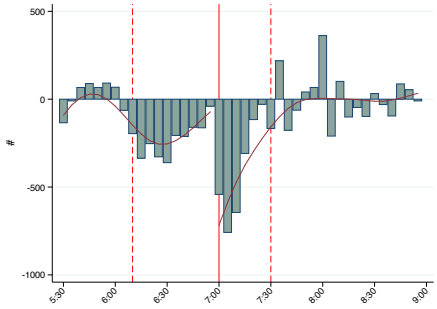


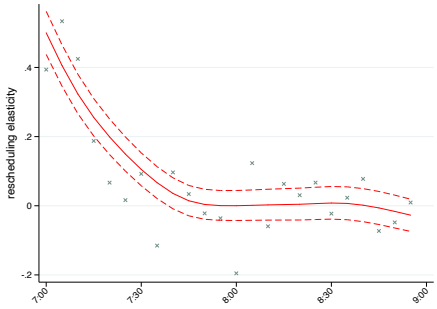
Rescheduling Elasticity



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

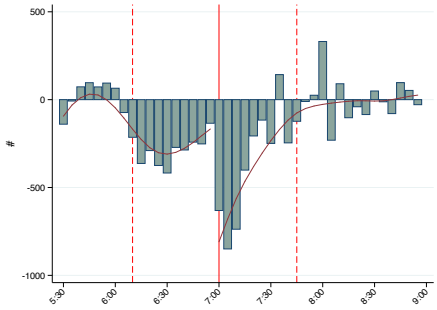
Panel E: Rescheduling Window between 6:15 AM and 7:30 AM

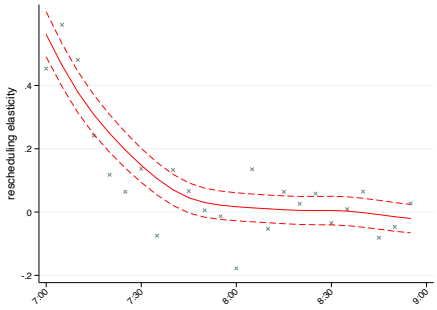




B=-2,059, M=2,446, Counterfactual number of trips in the missing window=28,882

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





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.

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.

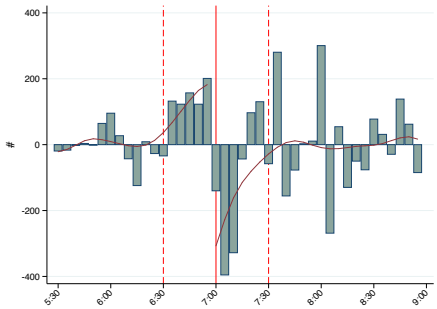
We first hold the bunching window (6:30 and 6:59 AM) unchanged but alter the missing window to be between 7 and 7:15 AM (Figure B.11 Panel A) and between 7 and 7:45 AM (Panel B). The missing and bunching masses are largely unchanged from the baseline. This is unsurprising because we have shown that the 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 substantially smaller than the missing mass. However, the counterfactual ridership curve is still robustly estimated, and the corresponding elasticity curves, defined for the time window after 7 AM, are little changed.

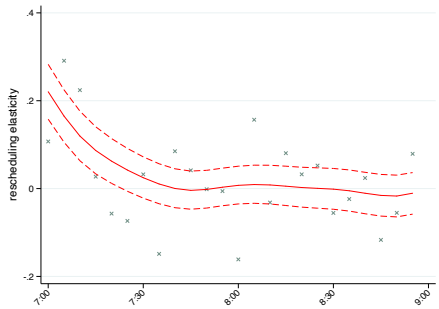
We expand the bunching window to between 6:15 and 6:59 AM in Panels E and F. While a wider bunching window allows for ample space for rescheduled trips to land, it leaves relatively few observations to the left of the bunching window for the estimation of the counterfactual curve. In cases of Panel E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

Figure B.12: 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 originate from one of the 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.

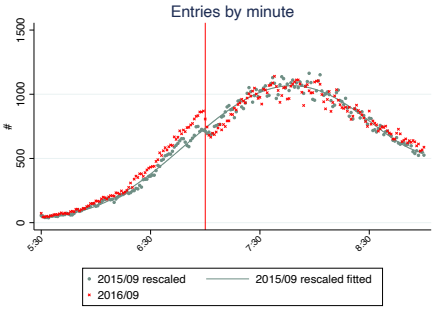
Joint estimation of demand and rescheduling elasticities. The baseline and robustness checks presented in Figure B.11 all impose a demand elasticity that is estimated from the OD pair regression discontinuity design. However, the demand elasticity that generates the number of opt-in trips due to the EBD may be different from the baseline. 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 of the elasticity, denoted as $e^{d,0}$. With the initial guess, we 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 $e^{d,1} = (B/M) \cdot e^{d,0}$, 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 B and M converge when e^d is -0.37, which is very close to the baseline demand elasticity. The converged value of missing and bunching masses converge at 1,172, which is also similar to that in the baseline (where M is 995, B is 1,263) and is small relative to the baseline ridership. Figure B.12 shows that the distributions of the bunching and missing masses (Panel A), as well as the rescheduling elasticities (Panel B), are also very similar to those found in the baseline.

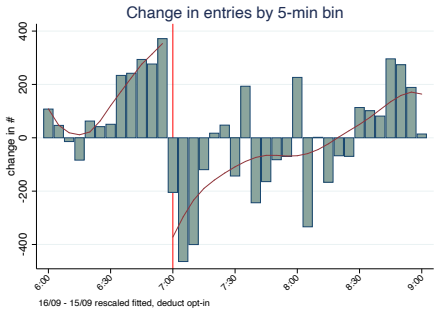
Figure B.13: Robustness Checks for Rescheduling Elasticity - Using September 2015 as Control

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

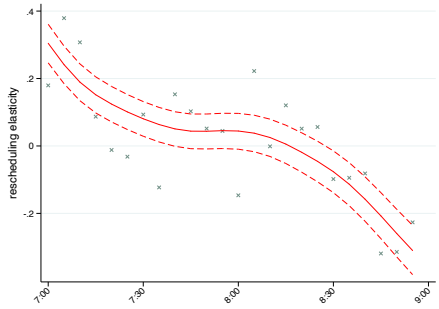


Panel B: Imposing Demand Elasticity

Missing Mass and Bunching



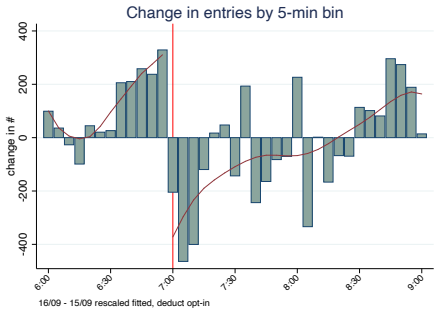
Rescheduling Elasticity



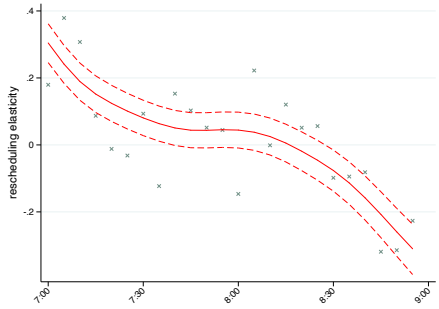
B=1,468, M=1,267, Counterfactual number of trips in the missing window=31,710.

Panel C: Joint Estimating 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.

An alternative approach is to find a control group that is not affected by the EBD. The ridership data from the week of September 15, 2015 serves as a natural comparison, as it was at the same time of the year but before EBD was implemented.

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 have increased between the two Septembers, we rescale the number of trips in September 2015 between 7:30 and 9 AM to match that in September 2016. The time window to calculate the scale effect is chosen to exclude the impacts of the opt-in and the 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).[6] The numbers of entries by minute in September 2016 are plotted in red dots in Panel A of Figure B.13, 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 exists 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 the 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 to estimate the demand and reschedule 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.13 plots the estimation results when the demand elasticity is imposed at -0.36. The graph 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 graph on the right plots the associated rescheduling elasticity. The rescheduling elasticity is around 0.3 for trips just to the right of the cutoff and declines quickly to

[6] 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.

zero by around 7:20 AM. Those estimates again correspond to a small share of peak-hour ridership 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 within the sample time window that is between 5:30 and 9 AM. 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 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 their overall usage as well as the 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 and has an area of 2,267 square kilometers) 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 directional. Commuters who travel from bin A to bin B in the morning and then from bin B to bin A in the afternoon are regarded to have trips in the same location bin pair, while those who travel from bin B to bin A in the morning and the opposite in the afternoon are regarded to have trips in a different location bin pair.

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

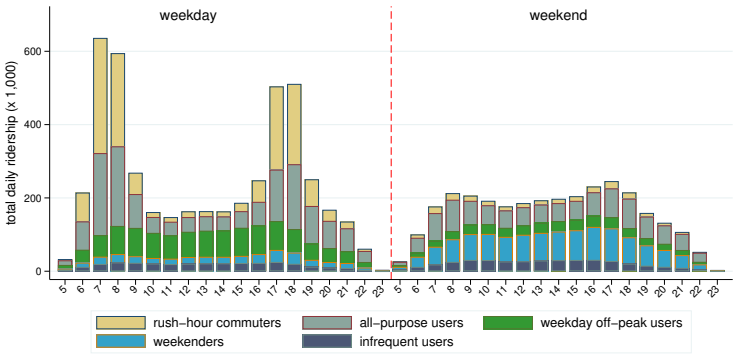
travel pattern is very regular. However, one may argue that her pattern is less regular than that of a user who takes ten trips, nine of which are in the same location bin pair, although the latter will have an HHI of less than one. The geographic concentration rate will say the former user has a value 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 bring the same amount of potential variation to the algorithm. The algorithm 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 a new group of users who exhibit travel patterns that are different from any of the other user groups.

Trip characteristics by user types are reported in Table 3. User types are mostly identified by the number and the timing of trips. We name each group with the most salient feature of their travel patterns: infrequent users, weekenders, weekday off-peak users, rush-hour commuters, and all-purpose users. The last two groups include most of the frequent users, regular commuters and all-purpose users are separated by whether their trips are concentrated in a handful location bin pairs.

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.

With users grouped into different types, Figure C.1 plots 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 AM in the morning

and between 5 and 7 PM 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. Rush-hour commuters and all-purpose users contribute to most of the ridership during peak hours. Weekday non-rush hour users 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 relatively evenly spread out across time and days.

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	all-purpose users	weekday off-peak users	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. Standard deviations are in parentheses.

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 (see Table 3). 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 rush-hour commuters who have regular commuting patterns; another one million cards are frequent users who travel to many destinations in both peak and off-peak hours; they are labeled as all-purposed users; about one million users travel mostly during non-peak hours on weekdays; and about 0.7 million users mostly travel on weekends. Table C.1 reports the characteristics of the trips by user groups.

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^{st}. However, a key component of the new fare structure is a cumulative quantity discount that depends on the 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 keep the randomly chosen 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. For example, if one user had two trips during the observed week while another had four trips, each of the four trips belonging to the latter user is twice as likely to be drawn than each of the two trips that belong to the former user. 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 the other weeks in September 2014, which are not included in our data. 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 has many drawbacks. For example, it inevitably misses many infrequent users who did not travel during the week for which we have data. To check whether the simulated monthly ridership makes sense, we take advantage of the fact that despite the comprehensive fare structure change and subway expansion,

the aggregate ridership and the distribution of users between September 2014 and April 2015 are largely comparable. In an average week in April 2015, there were about 28 million trips from 7 million unique users. In the week of September 16, 2014, the corresponding numbers were 29 million trips from 6.7 million unique users. In the full month of April 2015, there were a total of 124 million trips from 12.4 million unique users. The total number of trips is similar to that in the populated September 2014 data, and we know that infrequent users, though accounting for nearly half of the total users, make up for only 8% of the total trips (Table 3).

D.2 Algorithms to Calculate Ridership and Counterfactual Fares

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

Step 1. 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 distribution that is remarkably similar to those in April 2015. Appendix C.3 reports the details. The implicit assumption imposed here is the fare structure change does not systematically alter user types.

Step 2. We impose the demand elasticity specific to each user's type. Heterogeneous demand elasticities are reported in Figure 3. Also motivated by the pattern shown in Figure 3, we do not separate different types of trips within the same user type. For trips that are in the neighborhood of the 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 these two points. We assume the rescheduled trips evenly land between 6:30 and 6:59 AM.

Step 3. With demand and rescheduling elasticities at hand, we calculate the corresponding number of trips under the new fare structure for each trip in the simulated data that cover the full month of 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.

Step 4. We consider four different behavioral responses to cumulative quantity discounts. The oblivious, myopic, ironing, and rational users respond to the listing price, the instantaneous marginal price, the monthly average price, and the month-end final 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 expenditures and update

the corresponding prices. This procedure is repeated until the implied prices according to the assumed behavioral response are consistent with her choices 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 new 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 to the target.

D.3 Extrapolating Road Speed

To calculate the marginal external cost of traffic congestion (MECC, Equation 9), we need granular speed measures that vary by hour, day of the week, and location in the city. The main source of the speed measure comes from the replication data of Yang et al. (2020), which includes a random sample of road monitoring stations on Beijing's highways and ring roads in 2014. The monitoring stations are geocoded (location is relative to Beijing's five ring roads), report the hourly average speed of vehicles that pass the station, and record the volume of passing vehicles.

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). That data has 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.[7] The five ring roads cut Beijing into six regions ranked by the distance to the city center. We take the simple average speed of monitoring stations in the same area. We then weigh the average speed in each region 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 regions. We then sum up all trips to obtain the spatial distribution of subway trips. We calculate that about 22% of passenger-kilometers from

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

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. Both 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 kilometers per hour (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. The average hourly speed from Baidu Maps in the sample of road segments used by Gu et al. (2021b) is about 30 km/hr. Everything else equal, Equation 9 suggests that a higher speed leads to a lower MECC. Therefore, our calculations of congestion externality are likely underestimated.

Figure D.1: MECC by Hour

Note: Values of MECC are from Yang et al. (2020)

Figure D.1 plots the MECC as a function of hours separately for weekdays and weekends. On weekdays, the MECC exhibits a clear twin-peaked pattern, where the peaks correspond to morning and afternoon rush hours. Diverted trips from the subway during peak hours likely generate large congestion externalities. The MECC on weekends also exhibits a twin-peaked pattern, though they are less salient than that on weekdays and are lower.

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.

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 myopic users. Ridership and welfare outcomes under these two heuristically optimizing behavioral types are similar to each other 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 new fare structure. A revenue-equivalent flat rate results in slightly higher welfare losses but a somewhat smaller congestion externality. Together with Table D.1, those findings show that the new fare structure achieves better welfare and efficiency goals as long as users respond to the nonlinear budget at least heuristically. However, there is no empirical support for such quasi-optimal responses. On the other hand, a revenue-equivalent fare that adds a peak-hour premium to the new fare structure results in substantially larger welfare losses and congestion externality with myopic and ironing users.

Distributional welfare impacts under alternative fare structures. We consider welfare un-

der alternative fare structures separately for different user groups. We consider two polar behavioral types in response to the quantity discounts: oblivious and rational. Welfare calculations with ironing and myopic users in general lie between the two polar cases and are 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 under alternative fare structures. 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 (defined as the difference between the decline in consumer welfare and the increase in revenue), and (4) changes in the congestion externality. All measures are in terms of 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 regular commuters (Panel A), the new 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 new 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 new fare structure.

These findings are intuitive. The quantity discounts built into the new 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 discounted prices and take more trips, leading to lower losses in consumer welfare and a smaller congestion externality.

Compared with the new 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. Adding a peak-hour premium 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 (Figure D.1).

The welfare impacts for all-purpose users (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 they have a less elastic demand.

For less frequent users (weekday off-peak users, weekenders, and infrequent users, reported in Panels C through E), the alternative flat rate almost always performs better than the new fare structure. The loss in the consumer surplus, increase in the deadweight loss, and the additional congestion externality are mostly lower. Those patterns hold 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 when frequent users enjoy the discounts. 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. The fare level needs to be higher to generate the same revenue, and the impacts of higher fare levels disproportionately fall on 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 on weekends or during off-peak hours of the workdays. 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.2 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 for details). The graphs show the composition of trips from different user types, 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 commuters and all-purpose users account for the bulk of peak-hour trips. Ridership on weekends is substantially 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 that correspond to the temporary patterns of ridership under 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 another stated policy goal of the fare change that is not discussed in detail in our main analysis. That is, how different fare structures can pare peak demand during rush hours, easing 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 the peak-hour premium is the most effective in cutting peak load, regardless of behavioral assumptions. Peaking/Off-peak pricing mostly achieves that outcome by reducing trips by regular commuters and all-purpose users. However, due to the small rescheduling elasticity, 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.

Under the new fare structure, there are larger reductions in peak-hour demand if users are oblivious. This is because peak-hour trips are predominantly taken by frequent users who qualify for discounts. If they are rational, they will 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.2 illustrate the tradeoff between consumer welfare, congestion externality, and negative externality from transit system crowdedness.

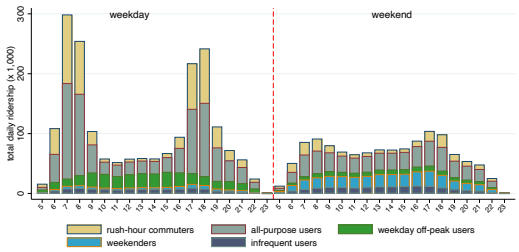
Table D.2: Ridership and Welfare under Alternative Fare Structures: Distributional Impacts

	Oblivious			Rational		
	current fare	alt. flat rate	peak/ off-peak	current fare	alt. flat rate	peak/ off-peak
	(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: All-purpose users (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 off-peak users (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]

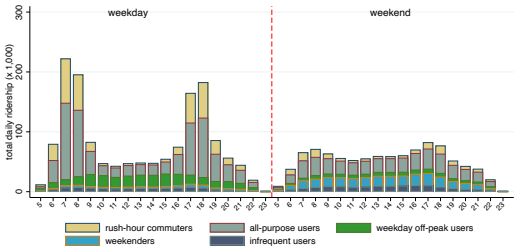
Note: This table summarizes ridership and associated welfare under alternative fare structures for different user groups. Counterfactual fare levels are calculated such that the aggregate revenue is the same as under the new fare structure with the specific behavioral responses to the quantity discounts.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

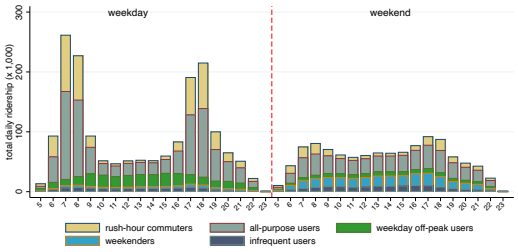
Figure D.2: Compositional and Temporal Travel Patterns under Alternative Fare Structures

Panel A: Under 2-yuan Flat Rate

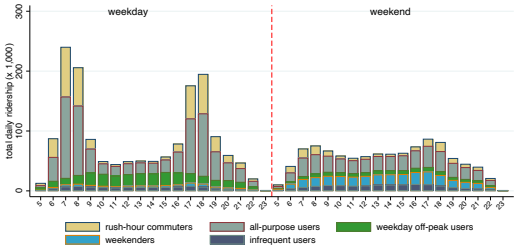


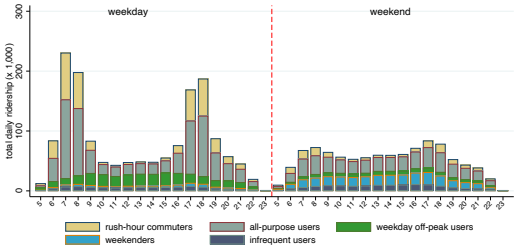
Panel B: Under the new Fare Structure



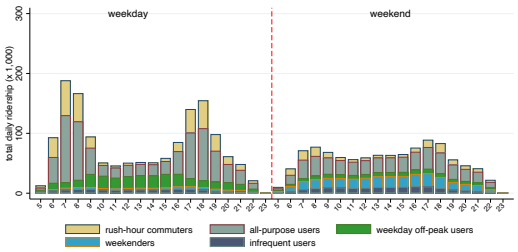


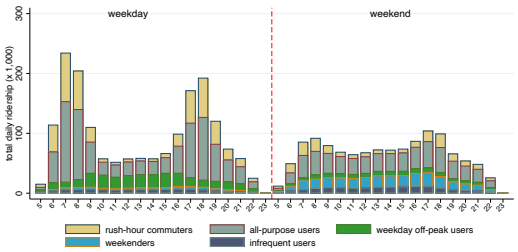
Panel C: Under the Alternative Flat Rate





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





Oblivious Rational

Note: The graphs plot the temporary ridership distribution from a 10% random sample of users from simulated full-month data covering September 2014. Graphs on the left target the aggregate revenue in the new fare structure with oblivious users. Graphs on the right assume rational users.

E Heterogeneity and Welfare Incidences by User Skills

Knowing how the fare structure change impacts users of different social economic statuses has important economic and policy implications. However, the subway ridership data does not have information on users' demographic and economic characteristics. In the main paper, we investigate heterogeneity and distributional welfare incidences on different types of users, where user types are categorized by their travel patterns. While those analyses yield interesting empirical patterns, they remain silent about whether the welfare impacts of the fare structure change were progressive or regressive.

Good data sources with demographic and income information at fine geographic levels are not available for Beijing.[8] In this appendix, we utilize a novel data set of commuting flows at fine geographic levels. The data include *imputed* skill shares of commuters, which we use to investigate the heterogeneity by user skill level. In the following, Section E.1 describes and validates the data; Section E.2 reports estimations of the demand elasticity by skill; Section E.3 reports the calculations the welfare impacts by skill.

E.1 Data Description and Validation

Sources of data. We use Baidu Map's grid-level commuting flows matrix that covers the entire Beijing. Commuting flows are imputed from location pins reported by Baidu Map's popular smartphone application or its location-service plugins used by many other applications. Specifically, Baidu receives regular, high-frequent location pins from those applications. Those pins allow Baidu to infer each device's "usual daytime location" and "usual nighttime location." Baidu maps call the former "work location" and the latter "home location." Aggregating home and work locations from millions of devices in Beijing, Baidu is able to calculate commuting flows at fine geographic levels. Chen et al. (2022) use similar data from Baidu to delineate commuting-based metropolitan areas in China, and provide detailed description of the technology and the data.

The commuting data used here is based on location pins in the three consecutive months ending in November 2017. The data we have access to does not include any device-level infor-

[8] One possibility is the population survey of 2015, which is a 2 in 1,000 random sample of the nationwide population. It includes basic demographic information about each individual and the worker's industry and occupation. Important for our purpose, it also has information on residence township, work township, and the usual mode of transportation for commuting. However, the sample size (with less than 20,000 workers living in urban Beijing, where the subway is relevant) is too small to credibly infer the demographic composition of commuters between all location pairs. Another possible data set is the Beijing Household Transportation Survey, which has about 50,000 households in the 2015 wave. The data set has basic demographics, home and work locations for workers, and annual household income. Furthermore, the location information is at the transportation analysis zone (TAZ) level, which is a smaller geographic unit than a township. The major downside of the data set remains its sample size. In addition, respondents were chosen using stratified sampling. Only a fraction of TAZs are selected, so the demographic and income characteristics associated with many station pairs cannot be matched.

mation. Instead, it is provided to us as a commuting matrix between home locations and work locations, where each location is a 700 meter-by-700 meter grid (each grid covers an area of 0.5 square kilometers). Beijing, with an area of more than 16,000 square kilometers, is divided into about 65,000 grids. Those grids are much finer geographic units than any available administrative definition. In contrast, Beijing consists of 16 districts, 334 subdistricts or townships, and about 7,100 communities.

We observe the number of commuters for each home-work grid pair. Notice that a "commuter" here corresponds to a device whose usual daytime and nighttime locations Baidu is able to capture reliably. Smartphones were ubiquitous in Beijing by 2017, and Baidu Maps was a popular application. The coverage is comprehensive, though not complete. The total number of commuters in the data amounts to some 9 million, while according to the statistical yearbook of Beijing, there were 12.5 million workers in the city by the end of 2017.[9]

Important for our purpose, the data provides numbers of commuters in three education levels. The first group corresponds to those with a high school diploma or less; the second group corresponds to those with a college or university diploma;[10] the third group corresponds to those with post-graduate degrees. Except for a small group of users who self-disclose their education levels to Baidu Maps, most users' education levels are not directly observed. Instead, education levels are *predicted* from various information about smartphone uses that Baidu Maps has access to. Such information includes self-disclosed demographic and social-economic characteristics when registering for Baidu Maps and Baidu's other applications,[11] home and work locations, places the user visits, applications installed on the device, and queries sent on associated applications. A machine learning algorithm is trained on those pieces of information to predict each user's education level. We do not have access to the underlying variables used for the prediction or the algorithm.

Comparison with Census 2020. We check the quality of the grid-level commuting data flows with the township-level tabulations of the 2020 Population Census. Beijing is divided into 334 townships. The census tabulations have each township's population by age and education. We aggregate the commuting flow data to the township level to facilitate comparison.

We conduct two sets of comparisons between the 2020 Census and the commuting flows matrix. The first comparison is of the number of workers by residence. The Census tabulation data does not report the number of workers, so we use the work-age (16-64) population instead. The second comparison is of the share of skilled workers (or work-age population) in each township, where skilled people are defined as those with a college degree or above, which corresponds to

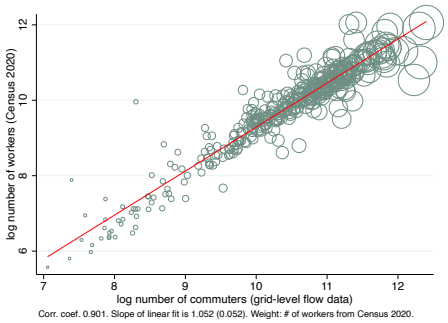
[9]Beijing Statistical Year Book 2018. URL: https://nj.tjj.beijing.gov.cn/nj/main/2018-tjnj/zk/indexch.htm. Last accessed in March 2024.

[10]A college diploma in China typically requires three years of study, while a university diploma requires four.

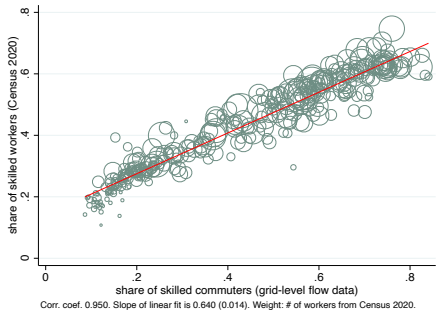
[11]Baidu also runs China's most popular search engine.

Figure E.1: Validation of Commuting Flows Data

Panel A: Number of Workers by Residence



Panel B: Share of Skilled Workers



Note: This figure compares the numbers and skill shares of workers from the grid-level commuting flows data and the 2020 population census township-level tabulations. The grid-level commuting flows data are aggregated at the level of residence township. Panel A shows the correlation between the number of commuters from the grid-level flows data (in horizontal axis) and the work-age population from Census 2020 (in the vertical axis). Panel B shows the correlation between shares of skilled workers (or work-age population for the census data). For the grid-level flows data, and the skilled ratio is defined as the share of commuters with a college degree or above. For the census data, it is defined as the share of the work-age population with a college degree or above. Each circle represents a township and the size of the circle is proportional to the township's population, measured using the census data.

education levels 2 and 3 in the commuting flows data.

Figure E.1 shows the correlations of the two measures. Each circle represents a township, and the size of the circle is proportional to its population. The red line is the linear fit of the circles, weighted by population size. In both cases, the measures from two data sources are highly correlated.[12] The correlation coefficient is 0.90 for the log number of workers by residence (Panel A) and 0.95 for the share of skilled workers (Panel B). The commuting flow data matches the official census data well at the fine geographic level.

The skill composition of subway riders. We use the commuting flows matrix to assign the skill composition of subway users between any pair of stations. We first define the "neighborhood" of a subway station as an area in which residents may potentially use the subway station as the origin or destination of their commutes. We define the a station's neighborhood the group of grids within a 2-kilometer radius from the station. Using those modes of transportation as feeders to the subway usually involves a short trip. For example, the median bus ride in 2015 was under 2 kilometers. Furthermore, using the Household Travel Survey of 2015, we find that trips that use the subway as the main mode of transportation typically involve an access time less than 15 minutes by walking, cycling, bus. Given those modes of transportation are slow, the access distance is typically less than 2 kilometers in straight-line distance. We then use the

[12]The levels are not identical because (1) the smartphone-based commuting flows data may not capture all workers, and (2) the definitions of workers in the two sources are different. For the census, we use the work-age population as a proxy for the number of workers.

commuter skill composition in the commuting flows matrix to generate the skill composition of commuters between pairs of neighborhoods. A grid can be assigned to multiple stations or none of the stations. We assume trips that start on weekday mornings are home-to-work trips, and those that start on weekday afternoons are returning trips.

For this approach to be valid, an implicit assumption is that high-skilled and low-skilled users in the same residence or work neighborhood have the same probability of taking the subway. To give an example in which this assumption is violated, consider the simple case where only the low-skilled take the subway while the high-skilled always drive to work; the skill ratio from the commuting flows data will misrepresent the skill share of subway riders.[13] Data from the 2015 Household Travel Survey shows a mild positive correlation between education level and the probability of using the subway for commuting.

This assumption can be tested. If high-skilled workers, conditional on their work-home location pair, are more likely to choose the subway for commuting, we should see a positive correlation between the skill share of commuters (the *skill composition*) and the share of commuters who use the subway (the *subway commute share*). The skill composition for each home-work neighborhood pair can be calculated from the smartphone-based commuting flows data. As a proxy for the subway commute share, we calculate the number of *trips* between the home-work neighborhood pair using the smartcard data. We use the number of trips instead of the number of unique users because a user can mix different modes of transportation for commuting. The log subway commute share is defined as

$$\ln\left(\frac{\text{\# of weekday trips}}{\text{\# of unique commuters}}\right)_{od},$$

where the subscript od indicates a home (o)-work (d) neighborhood pair.

Table E.1 Column 1 shows that the subway commute share is positively correlated with the skill share of commuters. This is consistent with the data from the Household Travel Survey, which shows that high-skilled workers are more likely to use the subway for commuting. Column 2 shows a positive correlation between the subway commute share and commuting distances. Intuitively, the subway has a comparative advantage in long-distance commutes. Both the share of skilled commuters and log distance are included as explanatory variables in Column 3, which shows that once conditional on the log distance between home and work locations, the skill composition of commuters is not correlated with the subway commuter share. Figure E.2 plots the residualized share of skilled commuters and the residualized log subway commute share. The scatter plot indicates a lack of visual pattern, and the linear fit is completely flat. Notice that when

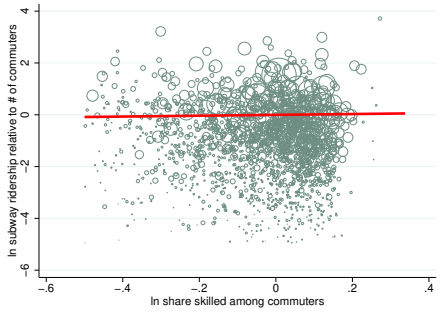
[13] However, the validity of the approach does not require the assumption that the probability of taking the subway is the same for any neighborhood pairs. It is allowed that people who live and work in nearby neighborhoods have a lower probability of taking the subway than those whose home and work locations are farther apart.

Table E.1: Skill Gradient of Transportation Mode Choice

	(1)	(2)	(3)	(4)	(5)
	log subway commute share			log share skilled	
log share skilled	1.743		-0.115		
	(0.078)		(0.071)		
log distance		0.968	0.977	0.076	0.198
		(0.015)	(0.016)	(0.001)	(0.001)
N	92,572	92,572	92,572	92,572	2,235,421

Note: In Columns 1 to 4, each observation is a home-work neighborhood pair. In Column 4, each observation is a grid pair (from the commuting flows matrix). In Columns 1 and 2, the dependent variable is the log subway commute share, which is defined as the log ratio between commuting trips between home station O to work station D to the number of commuters (from the commuting flows matrix) between the same pair. In Columns 3 and 4, the dependent variable is the share of skilled commuters. All regressions are weighted by the number of commuters in the pair. Robust standard errors are in parentheses.

Figure E.2: Mode Choice by Skill Shares



Note: The graph shows the relationship between the log subway commute share (subway ridership relative to the number of commuters) and the log share of skilled workers among commuters. Both variables are first regressed on the log distance between the home-work neighborhood pairs, and residuals are plotted. Each circle represents a home-work neighborhood pair, while the size of the circle corresponds to the number of commuters between the pair (from the commuting flows matrix).

we estimate the demand elasticity using the regression discontinuity design, a flexible polynomial of distance is always controlled for. So the endogenous mode choice of commuters is conditional orthogonal to the skill composition of users between a given home-work location pair.

The first three columns of Table E.1 suggest that the positive simple correlation between the skill composition of commuters and the subway commute share is solely driven by the fact that high-skilled workers tend to have longer commutes. This conjecture is confirmed in Column 4, which shows a strong positive correlation between the skill share and the commuting distance. In Column 5, we instead use the original grid-level commuting flows matrix, which shows a gradient of the skilled commuter share with regard to the log distance that is similar to the coefficient in Column 4. Those results lend credibility to the assumption that we can impute the skill com-

position of subway users from that among all commuters between the corresponding home-work location pairs.

E.2 Demand Elasticity by Skill

Table E.2: Heterogeneity in Demand Elasticity by Skill

	(1)	(2)	(3)
$\Delta \ln p_{od}$	-0.412	-0.458	
	(0.021)	(0.030)	
$\Delta \ln p_{od} \cdot SkillRt_{od}$		0.062	
		(0.029)	
$\Delta \ln p_{od} \cdot LowSkillRt_{od}$			-0.415
			(0.021)
$\Delta \ln p_{od} \cdot HighSkillRt_{od}$			-0.411
			(0.021)
N	52,957	52,957	52,957

Note: Each observation is a home-work location pair. Column 1 is estimated using Equation 1. Column 2 is estimated using Equation E.1. Column 3 is estimated using Equation E.1. In Column 3, the cutoff skill ratio is 0.69. The mean skill ratio among the above-median skill ratio OD pairs is 0.77, and that among the below-median skill ratio OD pairs is 0.63. Robust standard errors are in parentheses. All columns control for the polynomial of log distance up to the 5th order, the origin and destination fixed effects, and are weighted by the ridership in September 2014.

With subway users' skill composition assigned to each station pair, we estimate the heterogeneous demand elasticity with regard to skill. The estimation is based on the baseline specification, which exploits the price discontinuity due to abrupt cutoffs in the OD pair distance. Specifically, we estimate the following two versions of the OD pair RD model.

$$\Delta \ln(N_{od}) = \beta_1 \cdot \Delta \ln p_{od} + \beta_2 \cdot \Delta \ln p_{od} \cdot SkillRt_{od} + f(dist_{od}) + \Phi_{o,d} + \Delta \varepsilon_{od}. \tag{E.1}$$

$$\Delta \ln(N_{od}) = \gamma_1 \cdot \Delta \ln p_{od} \cdot LowSkillRt_{od} + \gamma_2 \cdot \Delta \ln p_{od} \cdot HighSkillRt_{od} + f(dist_{od}) + \Phi_{o,d} + \Delta \varepsilon_{od}. \tag{E.2}$$

Each observation in the regression is a home-work station pair. In Equation E.1, the change in log price is interacted with the skill share of commuters in the corresponding home-work neighborhood pair. The skill share ranges between 0 and 1, so β_1 indicates the demand elasticity of station pairs where none of the commuters are skilled; $\beta_1 + \beta_2$ represents the demand elasticity of station pairs where all commuters are skilled. We interpret β_1 as the demand elasticity of unskilled users, while $\beta_1 + \beta_2$ represents the demand elasticity of skilled users. For Equation E.1, we divide station pairs into halves by the associated skill share. The median skill share is 0.69.

The above-median station pairs have a mean skill share of 0.77, and the below-median station pairs have a mean skill share of 0.63.[14] γ_1 indicates the demand elasticity in OD pairs that have a skill share below the median, while γ_2 indicates that in OD pairs that have a skill share above the median. In both specifications, we use the *listing* price, given the overwhelming evidence that users do not respond to the discounts. We only include trips on workdays and assume all those trips are between home and work locations.

Table E.2 reports the results from those estimations. Column 1 replicates the baseline specification (Equation 1). The demand elasticity is -0.412. This is a larger (in magnitude) elasticity than the baseline because trips on weekends are not included here, and we know that weekenders, who make up a substantial share of the trips on weekends, have a relatively low demand elasticity. Column 2 shows the results from estimating Equation E.1. The demand elasticity of unskilled users is -0.458, while that of skilled users is slightly lower (in magnitude) at -0.396 (-0.458+0.062). Column 3 reports the results from estimating Equation E.2. The demand elasticity in more-skilled pairs is only marginally lower (in magnitude) than in less-skilled pairs. This is not surprising given (1) the distribution of skill rate across home-work pairs is rather tight, and (2) the difference in the demand elasticity of the skilled and the unskilled users is small (Column 2).

E.3 Welfare Impacts by Skill

Assigning skill to individual users Using the commuting flows matrix, we have determined the share of skilled commuters in any given pair of home-work stations. However, calculating the distributional welfare impacts of the fare structure change on users of different skills requires assigning a skill level to *individual* users. We use the following procedure to probabilistically assign the skill level to each individual user such that the aggregate skill share matches that of the population relevant to the subway – those who live *and* work in the neighborhoods of a subway station.

1. Trips on weekends and holidays are dropped. Those trips are less likely to be between home and work locations, so imputations of skill levels using commuting flows are less reliable. For trips on workdays, we assume those that start in the morning are from home to work, and those that start in the afternoon or evening are from work to home.

2. Each trip is assigned a probability of belonging to a skilled user. The probability is the same as the skill share of commuters between the home-work neighborhood pair. After drawing an independent random number, we assign the trip to a skilled user or an unskilled user.

[14] Beijing is one of the best-educated cities in China. The 2020 Census data show that the skilled share among the work-age population in the city was 52%. The subway network connects the urban part of the city. The skill share of the relevant population is even higher.

3. A user may have multiple trips and have trips between more than one home-work pair. So, some of the same user's trips may be classified as belonging to a skilled user, while others belong to an unskilled user. We take the average of those binary classifications within each user.

4. We decide a cutoff number, users with a share of trips classified as skilled above the cutoff number is classified as a skilled user. The cutoff number is determined such that the share of imputed skilled users matches the aggregate skilled share in all home-work location pairs from the commuting flows matrix data.

Out of the 6.7 million unique users in September 2014, we impute the skill levels of about 5.1 million users. The skills of users who only travel on weekends are not imputed. 69% of the users are skilled, matching the aggregate skill share of the subway-relevant population.

Welfare impacts by skill. Following the method described in Section 4, we calculate the welfare impacts of changing the 2-yuan flat rate to the new fare structure separately for the skilled and the unskilled. The demand elasticity for each group is from Column 2 of Table E.2 (-0.458 for the unskilled and -0.396 for the skilled).

Table E.3 reports the results of welfare calculations under various scenarios of behavioral responses to the non-linear monthly budget. Skilled users have a less elastic demand, which predicts a smaller change in ridership as the average fare increases. However, they also have longer trips (Table E.1), so they disproportionately experience larger increases in the listing price. Skilled workers also take more subway trips. Under the original 2-yuan flat rate, an average skilled user took 20.4 trips in a month, while an average unskilled user took 14.6 trips.

When users are oblivious to the quantity discounts, the reduction in ridership is steeper for the skilled users (-23% vs. -19.8%). When users respond to quantity discounts rationally, the reduction is similar (-11.3% for the skilled and -10.3% for the unskilled). This is because skilled workers take longer and more subway trips and benefit from the discounts more. Overall, the new fare structure is slightly progressive and the welfare incidence is largely similar for both skilled and unskilled users. If users are oblivious, the reduction in consumption surplus amounts to 92.4% of pre-fare rise expenditure for the skilled (37.8/40.8) and 87.3% for the unskilled (25.5/29.2). If users are rational, these two numbers are 85.3% (34.8/29.2) and 82.9% (24.2/29.2), respectively.

Table E.3: Aggregate Impacts of the Fare Structure Change

Panel A: Unskilled		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	14.6	11.7	-19.8%	13.1	12.2	12.2
Total distance (km)	206	157	-23.8%	176	164	165
Expenditure (yuan)						
before discount	29.2	51.6	76.7%	57.7	53.8	54.1
after discount	29.2	45.8	56.8%	49.1	47.0	47.2
Discount rate	-	2.1%	-	2.7%	2.3%	2.3%
Welfare impacts (yuan per user per month)						
Δ Consumer surplus	-	-25.5	-	-24.2	-24.7	-24.9
Δ Revenue	-	16.6	-	19.9	17.8	18.0
Δ DWL: -(ΔRev+ΔCS)	-	8.9	-	4.3	6.9	6.9
Δ Congestion externality	-	5.6	-	3.5	4.8	4.7
Panel B: Skilled		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	20.4	15.7	-23.0%	18.1	16.6	16.7
Total distance (km)	334	248	-25.7%	289	264	265
Expenditure (yuan)						
before discount	40.8	74.7	83.1%	86.6	79.2	79.7
after discount	40.8	63.5	55.6%	69.7	65.8	66.1
Discount rate	0	4.0%	-	5.1%	4.3%	4.4%
Welfare impacts (yuan per user per month)						
Δ Consumer surplus	-	-37.8	-	-34.8	-35.9	-36.4
Δ Revenue	-	22.7	-	28.9	25.0	25.3
Δ DWL: -(ΔRev+ΔCS)	-	15.1	-	5.9	10.9	11.1
Δ Congestion externality	-	10.0	-	5.2	8.2	8.0

Note: The table summarizes the welfare impacts of changing the fare structure from a 2-yuan flat rate to the new pricing schedule. Pre-fare-adjustment monthly ridership patterns are based on data from September 2014. Panel A reports the impacts on unskilled users; Panel B reports the impacts on skilled users. The user's skill is probabilistically inferred from the skill share of commuters in each home-work location pair. 69% of the users are skilled. The table reports consumer-level monthly averages.