How Taxi Drivers Terminate Their Shifts when Earnings Are Hard To Predict

Florian Artinger*[1], Gerd Gigerenzer†[1] and Perke Jacobs‡[1]

[1]Max Planck Institute for Human Development

Version 7, August 20, 2020

*artinger@mpib-berlin.mpg.de
†gigerenzer@mpib-berlin.mpg.de
‡jacobs@mpib-berlin.mpg.de

Abstract

Building on the assumption of rational expectations, the literature has modeled taxi drivers' shift termination choices using neoclassical and behaviorally informed versions of utility theory. We argue that rational expectations are insufficient to guarantee the rationality of utility maximization, which hinges also on the predictability of earnings. Using data of 11 million trips in Hamburg, Germany, we find minimal predictability of hourly earnings, leaving little reason to expect shift ends to be well described by utility maximization. A second analysis reveals that the overwhelming majority of shift ends is best predicted by satisficing models by which drivers terminate shifts when reaching an aspiration level on shift duration or earnings.

1 Introduction

In the late 1990s, Colin Camerer, Linda Babcock, Richard Thaler, and George Loewenstein [1997] proposed that taxi drivers conclude shifts after reaching their daily income target. This target-income hypothesis has initiated a fierce empirical debate, as it implies that increased earnings lead to shorter shifts. That is, if drivers are able to predict earnings in hours to come, they could profit from focusing work activity on profitable hours to realize their desired balance of leisure and income. Here, we argue that the empirical debate so far has neglected the perspective of the drivers and present two analyses that examine their informational constraints and behavior using a more stringent and flexible empirical approach than common in this literature.

We begin this article with a brief overview of the existing literature. The article by Camerer and colleagues presented an analysis of the wage elasticity of New York City cab drivers' labor supply. This elasticity describes the percentage response of working hours to percentage changes in wages, a parameter that is difficult to study in professions with fixed wages or salaries. The authors had determined that taxi drivers are useful to study because their hourly earnings fluctuate, implying that their shift earnings and lengths can be used to estimate their wage elasticity. Common sense suggests that this elasticity should be positive: with higher wages, drivers should be willing to work longer as the opportunity costs of leisure increase. Surprisingly, Camerer and colleagues found a negative elasticity across different samples of taxi drivers. In light of these findings, the authors speculated that drivers use a daily income target: If drivers terminated their shift as soon as they had earned a desired amount of income, this would imply that higher hourly earnings are associated with shorter shifts.

The income target was modeled as a discontinuity of the utility function. By this interpretation, drivers reach a decision by comparing the utility of terminating the shift with the expected utility of continuing the shift, and choosing the path yielding higher utility. Unlike neoclassical theory, the utility function includes a reference point at a particular income target. At this reference point, the marginal utility of income decreases so as to value additional income above the reference point less than below. Although the income target hypothesis is often formulated along the lines of the first interpretation, it is implemented along the lines of the second interpretation. For example, building on the seminal model of reference-dependent preferences by Köszegi and Rabin [2006], Crawford and Meng [2011], describe a utility function that includes both an income and a duration target. This model retains the neoclassical utility model and adds or subtracts utility based on whether income and working hours are below or above their reference points. In the remainder of this article, we will refer to such models as reference-utility models.

Both the negative wage elasticity and its explanation using reference-utility models have attracted considerable criticism. Most prominently, Henry Farber [2005] criticized the use of shift data and reported results from an analysis of trip data. Using a larger dataset of New York City cab drivers, Farber [2008] was able to estimate a model of shift termination that takes the end of each trip as a decision point whether to terminate a shift or not. He finds that the probability of shift termination increases discontinuously at a reference point but also notes that these reference points have little effect on decisions because they are either too high or too unstable. Following this example, subsequent analyses have derived and estimated shift termination models from neoclassical or reference-dependent utility theory. For example, Crawford and Meng [2011] estimated the parameters of their model using Farber's trip data. Because the neoclassical model is

nested within this model, the authors used statistical tests on their parameter estimates to conclude that their model offers better fit of the data than neoclassical theory.

Most recently, Farber [2015] used the full record of taxi trips in New York City over a period of five years to examine several aspects of the reference utility model. His lines of attack are threefold. First, he decomposes the variance in average hourly earnings to demonstrate that aggregate hourly earnings are, on average, possible to anticipate with sufficient precision. In models assuming that the reference point is set at the anticipated level of earnings, this implies that there is little room for reference dependence to play a role as it only affects deviations from the reference point. Second, he estimates a shift termination model and finds that termination probabilities are weakly related to previous earnings. Third, his wealth of data permits him to estimate elasticities for individual drivers and he finds them to be mostly positive. This collection of findings leads him to conclude that there is little evidence that income reference dependence is an important factor determining the labor supply of NYC taxi drivers.

Overall, the past twenty-five years of literature on taxi drivers' labor supply have focused on a specific paradigm, both theoretically and empirically. Theoretically, the literature has studied primarily the distinction between reference-dependent and neoclassical utility models. Presumably, this is due to several factors, including the discipline's focus on utility models and the analytical convenience of nested models. Empirically, the existing literature has started to complement aggregate estimations of the wage elasticity using shift data with estimations of shift termination models using trip data. However, the general approach has remained the same: An existing sample of data is fitted using a regression model and the confidence intervals of one or two variables are used to distinguish between competing behavioral models.

We argue that the finetuning of utility functions has been premature. To date, the literature has neglected the more fundamental question of whether utility theory provides the best framework to model taxi drivers' shift decisions. To shed light on this issue, we ask two questions and offer two analyses to answer these questions.

First, we ask whether we can expect drivers' choices to adhere to the predictions of utility theory. In the next section, we derive theoretically how the assumption of rational expectations is insufficient to imply rational choices under uncertainty, when agents are unaware of the mathematical structure of the decision environment. Under these circumstances, utility maximization hinges on accurate estimation of the relevant parameters. In the fourth section, we present details of data from more than 3,500 drivers in Hamburg, Germany. Using these data, the fourth section assesses the predictability of hourly earnings in the Hamburg taxi market. We first replicate Farber's earlier decomposition of hourly earnings and modify the analysis to better answer the question at hand. Next, we compare the predictive accuracies of various prediction models, including regularized regression. Across both analyses we find that the predictability of hourly earnings is minimal. Based on these results, we conclude that there is little reason to expect drivers' shift choices to be consistent with the predictions from utility maximization.

Building on this result, we ask whether drivers are better predicted by utility maximization or by satisficing heuristics. According to Herbert Simon [1955], satisficing heuristics use aspiration levels to make choices and do not require drivers to form expectations over future earnings. One of these heuristic models offers a more direct interpretation of the income target, hypothesized by Camerer et al. [1997]. By this model, drivers formulate a daily aspiration level on shift earnings and unequivocally terminate their shift after reaching the aspiration level. The fifth section reports a competitive test of six behavioral models for predicting drivers' shift ends. We find that the overwhelming

majority of shifts and drivers is best predicted by satisficing heuristics with aspiration levels on shift earnings or duration. In contrast, there are fewer than one percent of drivers are consistently best predicted by utility maximization. The final section discusses these findings and broader implications.

2 An Alternative to Rational Expectations

Taxi drivers face the problem of deciding when to terminate their shift. The most popular class of models for shift decisions is based on utility theory. According to these models, at specific decision points, agents compare the utilities of terminating and continuing the shift. Although utility functions vary, these models have in common that they assume that leisure and income are commensurable and agents are able to map their appreciation of these two goods on a continuous function. Using this function, both the utilities of terminating and continuing can be calculated.

Whereas the utility of terminating a shift can be computed from shift income earned so far and shift duration, the utility of continuing the shift is assumed to be based on expected income and expected shift duration. If drivers knew the probability distribution of their future earnings, the decision problem is one of risky choice and drivers could, in principle, use mathematical optimization to maximize their expected utility. Under these conditions, maximizing expected utility is rational and agents can be assumed to either rely on utility calculations or behave as if they did. By and large, however, taxi drivers lack a probabilistic description of their decision environment and face an uncertain rather than a risky decision environment [Knight, 1921].

Under uncertainty, forming expectations requires drivers to infer their decision environment. That is, drivers need to form their expectations based on a learning sample of simwilar decisions. Under these circumstances, agents are often assumed to form "rational expectations". John Muth [1961] coined the term, referring to agents using the relevant economic theory in forming expectations about macroeconomic variables. To date, the term has seen different definitions as the concept was applied to more diverse settings. In the case of taxi drivers, expectations are formed about individual drivers' earnings rather than macroeconomic outcomes, and are based on empirical induction rather than theoretical deduction. That is, drivers need to predict their future earnings rather than derive their value from a theory about supply and demand. We argue that for problems of this sort, rational expectations do not imply rational behavior. Unfolding this argument is somewhat complicated by the fact that there is no unified definition of rational expectations. We therefore consider two different definitions.

By one, rather stringent definition, expectations are rational if they are unbiased and based on all information [e.g., Samuelson and Nordhaus, 1998]. This definition does not imply that agents use the most accurate prediction possible. It is a widespread misconception that unbiased models predict well. To clarify this point, we include a brief decomposition of prediction error into bias and variance. This distinction is well-known in empirical sciences but notoriously under-appreciated in economic theory.

Suppose the task is to predict y, the value of an unseen item, based on its observables $\boldsymbol{x}$ and a model $m(\boldsymbol{x}, \boldsymbol{p})$ that is trained on a random sample of training data of size n. The error in prediction is measured by the root mean squared error, RSME, an can be decomposed into

$$\text{error} = b^2 + v + e, \tag{1}$$

where b denotes the bias in the predictions, v their variance, and e denotes the irreducible

error [Geman et al., 1992]. To understand bias and variance, recall that there exist L possible learning samples of size n, each of which yields its own predictions $\hat{y}_1, \hat{y}_i, ..., \hat{y}_L$. Bias is defined as

$$b^2 = \{E_n[\hat{y}_i] - E[y]\}^2, \tag{2}$$

where $E_n[\cdot]$ denotes the expectation with respect to different learning samples of size n and $E[\cdot]$ denotes the expectation with respect to unsystematic error. Bias refers to the difference between the average of the possible learnings samples and the true value and reflects a misspecification of the model. In contrast, variance is defined as

$$v = E_n[\{\hat{y}_i - E_n[\hat{y}_i]\}^2], \tag{3}$$

and refers to the variance of possible predictions around their expected value. Variance therefore reflects the model's sensitivity to idiosyncrasies in the learning sample. Artinger et al. [under review] offers a more detailed exposition.

Although bias and variance are influenced by multiple factors, they vary with the number of model parameters. In general, bias tends to decrease in the number of model parameters, whereas variance tends to increase in the number of model parameters [Hastie et al., 2001], creating a trade-off between bias and variance. This trade-off has important implications for model selection: Although unbiased models yield the best in-sample fit, unbiased models are not necessarily those yielding the best predictions. Indeed, the bias-variance trade-off implies that models seeking to reduce bias tend to incur excess error from variance. However, for prediction, the right balance of these two kinds of error depends on the particularities of the problem at hand. In many situations, using an unbiased model that uses all available information does not yield the best possible prediction and does not result in a rational expectation.

By a second, more generous definition, expectations are rational if "agents do the best they can with what they have" when it comes to the formation of expectations [Maddock and Carter, 1982, p.41]. This definition permits agents to use any model suitable for a given prediction problem. It assumes that agents are aware which model or algorithm yields the best balance of bias and variance and use that model. Defined in this way, rational expectations ensure that the agent uses the best prediction possible with the available information.

However, even in its generous definition, rational expectations do not imply rational choices. Rational expectations allow utility maximization to be based on the most precise predictions. However, it does not guarantee that utility maximization yields better choices than alternative decision strategies that do not require predictions. To see this, it is important to remember that under uncertainty, the agent does not know the probabilistic structure of the decision problem, but assumes a simplified structure and estimates the relevant parameters. Both these steps are necessary but also potential sources of errors.

First, the agent needs to simplify the existing structure of the decision problem — either purposefully to maintain computational tractability or inadvertently for a lack of knowledge. Any claim of optimality refers to the simplified problem. Depending on the nature and degree of simplification, the optimal choice in the simplified problem may differ considerably from the true best choice [Simon, 1979]. Second, the agent needs to estimate or predict the relevant parameters, such as expectations about future values. The quality of these estimates depends on the agent's choice of prediction model but also on the predictability of the decision environment. Even rational expectations can remain imprecise when randomness is high or available information is scarce or irrelevant.

Both the simplification of the problem and the need for parameter estimation imply that rational expectations are no sufficient condition for utility theory to yield rational

choices under uncertainty. In the case of shift decisions, forming expectations over future earnings is the Achilles heel of utility theory as well as any other theory requiring such expectations. Therefore, our first analysis examines the predictability of hourly earnings. If hourly earnings can be predicted with reasonable accuracy, this can strengthen the belief that utility maximization with rational expectations constitutes the normative benchmark for shift decisions. However, if earnings are difficult to predict, this casts doubt on the adequacy of utility maximization in this context.

An alternative approach to decision modeling are fast and frugal heuristics. The term heuristics borrows from the computer science literature, referring to an algorithm that ignores part of the available information to reach its goal. This ignorance can take many forms. Whereas heuristic algorithms in computer science can remain calculation-intensive [e.g., Pearl, 1984], heuristic decision models in cognitive science typically ignore large parts of the available information, resulting in strategies that are fast to implement and easy to communicate [e.g., Tversky and Kahneman, 1974, Gigerenzer and Goldstein, 1996].

Consider, for example, the task of predicting whether an existing customer will return for a purchase within a one-year time horizon. This classification task is common in marketing practice to target advertising efficiently. The task can be solved using the pareto-NBD model by Schmittlein et al. [1987], which uses the customer's full purchasing history to calculate the probability of a new purchase. In contrast, the hiatus heuristic requires only the time of the last purchase and predicts a return if that last purchase was after some threshold, and no return otherwise. This frugality makes the strategy easily applicable by marketing managers who seem to use it regularly. Wübben and Wangenheim [2008] find that this strategy can yield better decisions than the pareto-NBD, because it is less exposed to error from variance.

The hiatus heuristic is an example of the broader class of satisficing heuristics. The common denominator of these heuristics is their use of an aspiration level to reach a decision. An aspiration level is defined here as the threshold on one of the variables of interest that satisfies an aspiration and initiates an action [see also Lewin et al., 1944]. In the case of the hiatus heuristic, the aspiration level is the threshold separating returning customers from those who do not. Satisficing heuristics were fist described by Herbert Simon [1955, 1956] and later identified and studied across a range of decision tasks [for an overview, see Artinger et al., under review]. The earnings target described by Camerer et al. [1997] can be modeled more directly as a satisficing model.

In contrast to utility theory, satisficing heuristics do not require agents to form expectations about the future. We therefore expect these models to be descriptive of behavior when future hourly earnings are difficult to predict. In contrast, when hourly earnings are reasonably predictable, we would expect drivers to be better predicted by utility models. Our second analysis tests this hypothesis and compares the predictive powers of two different utility models and four different satisficing heuristics. Before both analyses are presented in turn, the following section presents an overview of the data used for analysis.

3 Hamburg Taxi Data

For the purpose of this study, we acquired data of taxi shifts and trips from Hamburg, Germany. The data used by Farber (2015) is unfortunately not openly available for the wider scientific community to analyze. These data comprise a sample of 6,998 drivers, 1,138,726 shifts, and 13,822,310 trips, collected electronically through so-called fiscal

taximeters. These devices are regular taximeters that use the cellular network to send trip and shift information to a secure server, where tax authorities can access the data and verify tax statements. The system is commercially maintained.

The data we obtained span the period from January 1, 2013 to December 31, 2015. During these three years, participation in the fiscal taximeter program was voluntary, but companies received free devices in exchange for participation. This incentive was substantial, as companies were aware that from 2017 the fiscal taximeter became mandatory nationwide. Indeed, by the end of 2015, two thirds of the 3,200 taxis in Hamburg used the fiscal taximeter [Levy, 2015]. Importantly for the purpose of this analysis, the sample is non-random as companies self-selected into data collection.

The data consist of two streams of information, one on shifts and one on trips. The shift stream presents a record of the beginning and end of a shift, as drivers log in and out of the taximeter. Thus, a shift can mean any period of time that the driver defines as such. The shift stream is independent of the trip stream, which keeps a record of each trip and its associated data. In Hamburg, trips can result from taxis being hailed on the street, drivers waiting in lines on places of high demand, drivers accepting trips through smartphone apps, or drivers cooperating with telephone centers that allocate trips among members. Together, both streams give an overview of the entire shift period from the start of the shift, to the first trip, to the last trip, and the point in time when the driver concluded the shift. For both shifts and trips we observed the variables described in Table 1.

In addition to the taximeter data, we collected data on observables that we suspected may affect daily demand. First, we identified days with events that may have caused surges in the demand for taxis. These were public holidays (41 days in the observation window), soccer games with Germany playing at the 2014 FIFA world cup (14 days), Hamburg's annual harbor fair, the biggest in the city (12 days), strikes in public transu8portation (17 days), strikes at Hamburg airport (10 days), and school vacations in Hamburg (157 days). Second, we obtained weather data from the Hamburg weather station, recording the amount of rain per square meter, both per day [DWD Climate Data Center, 2018a] and per clock hour [DWD Climate Data Center, 2018b]. These data were matched with each shift based on the day of shift begin and with each trip based on the clock hour of trip end. An overview of the additional variables used in the analyses is given in Table 1.

The separation of shift and trip data allowed us to check their consistency. In a first step, we checked each stream for internal plausibility, for example, whether trip begin is before trip end. In a second step, we combined the two data streams and checked whether the cumulative trip data is consistent with the recorded shift data. For example, we tested whether the first and last trip associated with a given shift fall within the period between shift begin and shift end or whether the shift totals of kilometers or earnings matched the cumulative trip data. When we determined an inconsistency, we deleted the corresponding shift and all associated trips from the dataset, unless we deemed the error to be small and fixable. This happened in either of three cases. First, when the shift data gives the correct trip count and all trips took place between shift begin and end, but the cumulative earnings does not match. In this case, shift earnings was set to the cumulative trip total. Similarly, when the cumulative trip earnings matched shift earnings and all trips took place within the shift period, but the number of trips in the shift data was incorrect. In this case, we set the trip count in the shift data equal to the number of trips observed in the trip data. Finally, when all trips fall within the shift period but the trip count in the shift data is off by one. In this case, we adjusted the trip

Table 1
Selected Variables Used in Analyses

	variable	description	source
shift data only	`sbeg`	shift begin	taxi data
	`send`	shift end	taxi data
	`searn`	shift earnings in EUR	taxi data
	`sdur`	shift duration in minutes	taxi data
	`precip.d`	daily precipitation per square meter	weather record
trip data only	`tid`	trip ID	taxi data
	`tbeg`	trip begin	taxi data
	`tend`	trip end	taxi data
	`tearn`	trip earnings in EUR	taxi data
	`tdur`	trip duration in minutes	taxi data
	`precip.h`	hourly precipitation per square meter	weather record
data in common	`did`	driver ID	taxi data
	`sid`	shift ID	taxi data
	`stype`	shift type	taxi data
	`ctype`	company type	taxi data
	`new`	day after fare increase with minimum wage	public record
	`weekday`	day of the week	public record
	`worldcup`	day with Germany playing world cup (shock)	public record
	`harbor`	day during annual harbor fair (shock)	public record
	`strike`	day with strike in public transport (shock)	public record
	`airport`	day with strike at Hamburg Airport (shock)	public record
	`holiday`	public holiday (shock)	public record
	`vacation`	day during Hamburg school vacations (shock)	public record

count to the number of trips observed in the trip data. In total, we made adjustments to 327,320 shifts with 5,838,654 trips. In all other cases of inconsistency, we deleted the corresponding shift and all associated trips from our data. This happened with 70,178 shifts with 937,340 trips.

Two events fall into our observation window that have plausibly changed market incentives. First, on September 16, 2014, the Hamburg Senate raised the taxi fare by about eight percent on average[1], effective from October 1, 2014. Similarly to other German cities, taxi fares in Hamburg are regulated by the local authority and reviewed every few years. The 2014 increase took place in anticipation of the second major change in the taxi market, the introduction of a national minimum wage in Germany. In August of 2014, the German government introduced a nationwide minimum wage of 8.50 EUR per hour, effective from January 1, 2015. This minimum wage was long expected and relevant to German taxi markets as many drivers had earned below the minimum wage. The new minimum wage applied to all employed drivers but not to self-employed drivers. Taxi companies with employees responded to the minimum wage in different ways, from incentives to maximize revenue to regulations that declared waiting periods as stand-by time. In addition, some companies were split up, making drivers henceforth self-employed. Because the two changes are likely to effect taxi drivers' behavior, we have split the observation window into two parts. The *old market* lasts from January 2013 to September 2014 and is characterized by pre-increase fares and no minimum wage. The *new market* lasts from January 2015 to December 2015 and is governed by increased fares and a minimum wage. Shifts and trips in the three months between these two periods

[1]Specifically, the base charge increased from 2.90 EUR to 3.20 EUR, the rate for the first four kilometers increased from 2.20 EUR to 2.35 EUR, for the next five kilometers from 1.90 EUR to 2.10 EUR and afterwards from 1.40 EUR to 1.45 EUR.

were ignored.

We classified shifts as day and night shifts to examine them separately where necessary. Specifically, shifts starting before noon are classified as day shifts, whereas those starting after noon are classified as night shifts. These two shift types have distinct demand profiles. Aggregated across all days, peaks in demand occur between 7am and 11am, as well as between 4pm and 10pm. Whereas day shifts tend to cover the morning peak, night shifts tend to cover the evening peak. Because these peaks occur at different points during the shifts[2], we decided to examine driver behavior separately for each of them.

We also placed additional restrictions on the data to receive our final dataset. First, we restricted shifts to those lasting shorter than 24 hours. This way, we exclude drivers that own their car and work long hours with many mid-sized breaks. Second, we excluded shifts with fewer than three trips to exclude shifts that were likely concluded prematurely. Third, we removed shifts that started on days where either the observation window started or ended (January 1, 2013 and December 31, 2015) to be sure that shift data is complete. Fourth, we removed shifts taking place on days with switches to and from daylight saving time, concerning six days during the three-year period. Finally, we restricted our analysis to drivers with more than 25 day shifts or 25 night shifts or both, to have enough material for out-of-sample prediction. The final data set consisted of 3,408 drivers with 866,341 shifts and 10,966,838 trips.

Finally, Hamburg taxi drivers work at one of three types of companies: Those with multiple cars and multiple drivers, those with one car but multiple drivers, and those with one car and one driver only. We refer to drivers in the last category as single drivers, as these drivers own their taxi and are not subject to the minimum wage legislation or the restrictions of a shift schedule. The working arrangements of these drivers resemble the closest to those of New York drivers. In total, the data comprise 807 single drivers with 285,226 shifts and 3,245,219 trips. Where appropriate, we examine single drivers separately to see whether findings are an artifact of employed drivers' working conditions.

4 Can Drivers Predict Next Hour's Earnings?

The first of our two analyses is an empirical examination of the predictability of hourly earnings. In one form or another, each utility model assumes that drivers compare the utility of terminating with an expected utility of not terminating their shift. That is, these models assume that transitory wage variation can be predicted with sufficient precision, such that drivers can form expectations over their utility if they were to continue driving. We refer to this assumption as the predictability assumption.

The predictability assumption underlies the normative assertion that labor supply elasticities be positive. Suppose hourly earnings would not vary during the day. Under such a regime, drivers could choose their working hours as soon as they learn the hourly earnings for the day. However, when hourly earnings fluctuate during the day, drivers need to decide incrementally whether the next hour will be worth their time. Whether or not drivers are able to concentrate their working time on profitable hours then depends on the predictability of these earnings: If hourly earnings are known with certainty in advance of each hour, drivers can, in principle, compare them with past earnings

[2]Figure A1 in the Appendix A shows both the number of shift and trip begins across clock hours, separately for day and night shifts. We observe that in day shifts, most drivers appear to start their shifts around the time of the first peak, whereas for night shifts, most drivers start considerably before the evening peak.

and decide whether the shift is worth extending. However, if hourly earnings are fully random, drivers trying to shift working hours to profitable periods find themselves unable to do so. Therefore, the normative assertion that elasticities be positive requires that hourly earnings can be predicted with a sufficient level of precision.

Farber [2015] has presented a systematic analysis of the predictability of hourly income using his data from New York City. To this end, he calculated for each of the 43,824 clock hours in his observation window the hourly earnings, averaged across all of the 8,802 drivers on duty during that hour. He then used two OLS regressions to relate the variance in average hourly earnings to variation in variables that are readily observable. The first of these models regressed average hourly earnings on a dummy for each year and a dummy for hours after the fare increase. The variance of the predicted values of this model represents permanent wage variation. The second model regressed the residuals of the first model on a dummy for each week of the year, a dummy for each hour of the week, and a dummy for public holidays. He refers to the variance explained by the second model as transitory but anticipated, whereas the variance left unexplained by both models is transitory and unanticipated. He finds that almost ninety percent of the variance in average hourly earnings can be anticipated, either because it is permanent or transitory but anticipated. He concludes that drivers can predict hourly earnings with a sufficient level of precision.

We argue that this conclusion is unwarranted for two reasons. First, it confuses the aggregate level of analysis with the individual level. From predictable aggregate earnings we cannot conclude predictable individual earnings, as the former are necessarily easier to predict. Indeed, insurances are based on this discrepancy between the predictability of aggregate and individual outcomes. Second, the in-sample fit of a statistical model yields limited conclusions about the ability to predict accurately outside of the sample or population. Because drivers can calibrate their regression model only on past data, the model's learning sample may be systematically different from the sample to which it is applied. Accordingly, the accuracy in out-of-sample prediction can be substantially lower than in in-sample fitting. These two shortcomings call for a more detailed examination of the predictability of hourly earnings. We therefore replicate Farber's original analysis and then extend it by an analysis of drivers' ability to predict their hourly earnings.

4.1 Variance Explained in Fitting

To replicate the original analysis by Farber [2015], we calculate $\log[\texttt{earn}_{h,i}]$, the natural logarithm of earnings of individual i in clock hour h. We then aggregate across drivers to obtain $\log[\texttt{earn}_h]$, denoting the natural logarithm of earnings during clock hour h averaged across all drivers active during that clock hour. This dependent variable is then modeled as follows

$$\log[\texttt{earn}_h] = \alpha_0 + \alpha_1 \texttt{y}_h + \alpha_2 \texttt{new}_h + \epsilon_h \tag{4}$$

where $\texttt{y}_h$ denotes a dummy for the year and $\texttt{new}_h$ denotes a dummy for the new market conditions with increased fares and minimum wage. The variance of its predicted values $\log[\widehat{\texttt{earn}}_h]$ represents the portion of variance in $\log[\texttt{earn}_h]$ explained by permanent changes in demand. In contrast, the variance of the residuals ϵ_h represent the portion of $\log[\texttt{earn}_h]$ unexplained by permanent changes. These residuals are then modeled as follows,

$$\epsilon_h = \beta_0 + \boldsymbol{\beta}_1 \texttt{w}_h + \boldsymbol{\beta}_2 \texttt{dh}_h + \beta_3 \texttt{holiday}_h + \gamma_h \tag{5}$$

where $\texttt{w}_h$ denotes a vector with 51 dummies for the week of the year, $\texttt{dh}_h$ denotes a vector with 167 dummies for the hour of the week, and $\texttt{holiday}_h$ denotes a dummy for a public

Table 2
Hourly Earnings: Variance Explained in Fitting

			Percent of Total Variance		
				Transitory	
Minimum Number of Drivers per Hour	Hours	Total Variance	Permanent	Anticipated	Unanticipated
1	15,106	0.02	6.29	61.70	32.01
25	15,022	0.02	6.31	62.71	30.98
50	14,095	0.02	6.06	65.20	28.74
75	13,340	0.02	5.76	67.26	26.98
100	12,576	0.02	5.52	69.40	25.08
125	11,642	0.02	4.99	71.82	23.19
150	10,412	0.02	4.60	74.29	21.12
175	9,194	0.02	4.31	76.62	19.06
200	8,252	0.03	4.33	78.29	17.38
225	7,300	0.03	3.78	80.56	15.66
250	6,340	0.03	3.40	82.65	13.94
275	5,321	0.03	2.87	85.09	12.04
300	4,356	0.03	2.36	86.63	11.01
325	3,568	0.03	2.13	87.61	10.26
350	2,906	0.03	1.81	88.59	9.60
375	2,224	0.03	1.93	89.02	9.05
400	1,570	0.03	1.01	90.15	8.83
individual drivers	3,437,503	0.70	0.04	4.30	95.66

holiday. The variance in the predicted values of this regression represent transitory but anticipated variation in $\log[\texttt{earn}_h]$, whereas the variance of the residual γ_h represents transitory and unanticipated variation.

For this analysis, we use data from 2014 and 2015 only. Of the $2 \times 365 \times 24 = 17{,}520$ clock hours between January 1, 2014 to December 31, 2015, there were 15,106 clock hours for which there was at least one driver active so that we could calculate average earnings. During the remaining clock hours, no driver in our data was on shift. In a first step, we estimate equations (4) and (5) using all available clock hours, yielding the variance decomposition shown in the first line of Table 2. In this analysis, the total variance in $\log[\texttt{earn}_h]$ to be explained is 0.023, and around 6 percent of this variance is due to permanent changes, whereas around 62 percent is due to anticipated transitory changes. This leaves 32 percent of the variance unexplained, about three times the share found by Farber [2015].

To understand this result, recall that Farber's data comprise 8,802 drivers, considerably more drivers than ours — on the one hand because New York City is much larger than Hamburg and on the other hand because our data is only a subsample of all drivers in Hamburg. For this reason, hourly earnings are averaged over fewer drivers. In some clock hours, there is only one driver active. We therefore repeated the analysis, considering only clock hours with a minimum number of active drivers. The results of these analyses are shown from the second line of Table 2 onwards. As the required minimum number of drivers per hour increases, we find that the unanticipated portion of the variance decreases. For example, when considering only clock hours with at least 300 drivers active, we have 4,356 hours left for analysis with a total variance of 0.031. Of these hours, the analysis above leaves around 87 percent unexplained, similar to the findings reported by Farber [2015]. We repeated the analysis with additional variables on demand shocks, but the results remained virtually identical.

The change in the proportion of unexplained variance illustrates our argument that high shares of explained variance result from high levels of aggregation. Such aggregation is useful to estimate overall demand or some measure of aggregate behavior. However, we argue that an assessment of individual drivers' potential to predict future earnings, necessarily has to examine the decision environment of individual drivers. With an average of 1.5 to 2.1 trips per clock hour, drivers' hourly earnings depend crucially on the profitability of individual trips. To illustrate, consider three passengers reaching Hamburg central station at 9.24am on the same train, two of which need to get to hotels around the corner and one needs to get to the airport outside of the city. The exact assignment of passengers to the first three taxis in line outside of the station determines which driver is looking at a profitable 30-40-minutes-trip to the airport and which driver spends five to ten minutes on a short trip and gets back to the end of the taxi line. Although consequential, this assignment is random in the sense that it depends on a plethora of unknown factors, including which passenger exits the train closer to the escalator or walks faster. One may argue that in the face of such randomness, the aggregate pattern is the best indicator drivers have. Although this may be true, it does not imply that a predictable aggregate pattern is useful for individual drivers. This depends on the magnitude of the pattern relative to the magnitude of random noise.

To obtain a better picture of the predictability of individual drivers' hourly earnings, we repeat the variance decomposition once more without any aggregation across drivers. Instead of $\log[\texttt{earn}_h]$, the log earnings averaged across drivers, we use the original variable $\log[\texttt{earn}_{h,i}]$, the log earnings of an individual driver. This analysis comprises earnings of 3,437,503 clock hours of individual drivers, which are modeled as follows, with

$$\log[\texttt{earn}_{h,i}] = \alpha_{10} + \boldsymbol{\alpha}_{11}\texttt{y}_{h,i} + \alpha_{12}\texttt{new}_{h,i} + \epsilon_{h,i} \tag{6}$$

modeling permanent variation, and

$$\begin{aligned}\epsilon_{h,i} = \beta_{10} + \boldsymbol{\beta}_{11}\texttt{w}_{h,i} + \boldsymbol{\beta}_{12}\texttt{dh}_{h,i} + \beta_{13}\texttt{holiday}_{h,i} + \beta_{14}\texttt{vacation}_{h,i} \\ + \beta_{15}\texttt{worldcup}_{h,i} + \beta_{16}\texttt{harbor}_{h,i} + \beta_{17}\texttt{airport}_{h,i} + \beta_{18}\texttt{strike}_{h,i} + \gamma_{h,i}\end{aligned} \tag{7}$$

modeling anticipated transitory variation, where subscript i denotes the driver and subscript h denotes clock hour, as before. In addition, we have added all additional dummies of demand shocks listed in Table 1.

The results of this decomposition are shown at the bottom of Table 2. As expected, the total variance to be explained is considerably larger than the variance of average hourly earnings and the portion to be explained by the models in equations (6) and (7) is much smaller: Jointly, only 4 percent of the variance is either due to permanent or anticipated transitory changes. By implication, 96 percent of the variance is transitory and unanticipated. This contrasts starkly with the results obtained earlier for the aggregate level and appears to be a direct consequence of the level of analysis.

4.2 Error in Prediction

As a next step, we recognize that drivers need to generalize from the past to the future. So far, we have examined predictability by explained variance in fitting the full set of observations. When deciding whether to continue her shift, a driver does not have access to future data points — but needs to predict outside of her learning sample[3]. This analysis

[3]In fact, one may argue that drivers need to predict outside of their learning population, depending on the stationarity assumptions one is willing to impose on the demand function. Consequently, we assume stationarity, such that drivers need to predict out-of-sample.

Algorithm 1
Competitive Test of Models for Predicting Next-Hour Earnings

```
foreach trip t do
    1. select random learning sample of 1,000 trips ending before start of t, irrespective of driver;
    2. record observed next-hour earnings after t;
    foreach model m do
        3. calibrate the model on learning sample;
        4. use necessary covariates to calculate predicted next-hour earnings for t;
        5. calculate residual between predicted and observed next-hour earnings;
    end
end
6. calculate for each model the root-mean squared residual across all trips.
```

mirrors this setup and examines the accuracy in predicting future earnings after each trip, based on a hypothetical learning sample of past trips.

To this end, we extract all trips that took place (a) after January 1, 2014, (b) at least one hour after shift begin, and (c) at least one hour before shift end. For each of these 6,149,686 trips t, taken by driver i, we sum up the earnings of all trips by i finishing in the 60 minutes after t was completed. This quantity gives the variable of interest and we refer to it as next-hour earnings. Note that this approach is not identical to calculating earnings across clock hours, as more than one trip can end per clock hour. Instead, we follow Farber [2005] in assuming that drivers use trip ends rather than full clock hours as decision points for terminating or continuing shifts and try to predict earnings during the subsequent 60 minutes. To obtain the learning sample used for prediction, we create for each trip t a random sample of 1,000 trips taken by any driver in the 180 days prior to t[4]. This sample is specific to each trip t. The learning sample includes trips of other drivers to reflect exchanges among colleagues about demand on different days. For each trip in the learning sample, i is assumed to be aware of realized next-hour earnings and the following covariates: earnings in the hour before, average next-hour earnings after all previous trips of i, current trip number, current shift earnings, current shift duration, a dummy for the year, dummies for the week of the year, dummies for the hour of the week, a dummy for a rainy hour, and dummies for the demand shocks listed in Table 1. In this way, we obtain 6,149,686 trips and their associated learning samples.

For this analysis, we adopt the perspective of the driver rather than the omniscient analyst. It differs from the previous analysis in two respects. First, it uses a limited sample of past trips to make predictions over future earnings. The error in predicting outside of the learning sample is typically larger than the error in fitting, particularly when the learning sample is small. Second, this analysis extends the number of cues that are available to each driver. Whereas the variables in the previous analysis reflected permanent changes as well as demand cycles and shocks, we supplement these data with variables about specifics about the current shift, such as cumulative shift earnings and the driver's average hourly earnings in the past. With the inclusion of additional variables on the one hand and the change in methodology on the other, it remains unclear how well hourly earnings can be predicted.

As explained in section 2, accurate prediction requires a good balance of bias and variance. Because the model with the best balance cannot be determined a priori, we test different candidate models competitively by feeding each of them with the same data and comparing the accuracy of their predictions. This approach is common in machine

[4]Indeed, we do this separately for trips before and after the fare increase. That is, for trips after the fare increase, the learning sample is restricted to other trips after the fare increase.

learning applications where models typically span different, non-nested classes of models, from linear models to non-linear ones such as decision trees or neural networks. The procedure of this analysis is summarized by Algorithm 1.

Two of the five candidate models are regression models. The first candidate model, REG, is elastic-net regression of all covariates listed above [Zou and Hastie, 2005]. The model is similar to OLS regression but addresses the exposure of OLS to error from variance by means of regularization. Regularization refers to a penalty of model complexity by "shrinking" the estimates, that is by subtracting a function of the OLS coefficients from the estimated value. Compared with OLS regression, the elastic net model has two additional parameters, λ and α, determining the amount and kind of shrinkage, respectively. In the study at hand, these parameters are determined based on 10-fold cross-validation from all data, where $\lambda = 0.013$ and $\alpha = 0.12$ were found to yield the best performance. Because elastic net regression reduces to OLS when $\lambda = 0$, the model used here yields predictions very similar to OLS. The second candidate model we use, LOG, is similar to model REG but first transforms all continuous variables into their logarithmic versions and re-transforms the predicted values back to EUR.

Apart from these computationally intensive models, we also test three simpler models that ignore part of the available data. The final three candidate models capitalize on low variance and each use only one variable to predict earnings in the following hour. Specifically, candidate model PAST uses the driver's the average next-hour earnings of *all past* trips to predict earnings in the next hour. In contrast, candidate model LAST uses earnings of only the *previous* hour to predict earnings in the next hour. Finally, candidate model MEAN does not use any covariate but the average next-hour earnings in the learning sample to predict earnings. In contrast to the regression models above, these models can, in principle, be implemented by drivers using no more than pen and paper and basic arithmetic.

Table 3 shows the results. Across all 6,148,059 trips, next-hour earnings to be predicted are on average 16.85 EUR and vary with a standard deviation of 13.87 EUR. The standard deviation can be viewed as the root mean squared error of the mean: If the average across all trips was known and used to predict next-hour earnings for all trips, the RMSE of these predictions would be equal to the standard deviation. For this reason, we use the standard deviation as a benchmark for the RMSE of other prediction models.

Consider first the two regression models, REG and LOG, shown in columns six and seven of Table 3. For the REG model, the root mean squared error in prediction is 12.42 EUR. Compared with the standard deviation, this equals a 10 percent reduction in the error. In contrast, the LOG model yields a RMSE of 14.39 EUR, considerably higher than the REG model.

Consider next the three simple models in columns eight to ten of Table 3. Whereas model LAST yields errors higher than the standard deviation, models PAST and MEAN yield predictions similar in quality to elastic-net regression with the errors of the PAST and REG models being almost indistinguishable. This finding offers a practical illustration of the bias-variance trade-off: Despite the fact that the REG model relies on all of the available data, its predictions are, on average, hardly better than those of the model that ignores most of this information. This result indicates that the additional variables, combined with the linear structure of the regression model expose it to higher variance in its predictions.

Irrespective of the model, the findings are sobering. Even the best model yields predictions that are only slightly better than the standard deviation of next-hour earnings. This finding implies that the models tested here are not fit to yield useful predictions

Table 3
Hourly Earnings: Error in Prediction in EUR

	Next Hour's Earnings			Model RMSE				
Trips	Median	Mean	SD	REG	LOG	PAST	LAST	MEAN
6,148,059	15.90	16.85	13.87	12.42	14.39	12.58	17.25	13.86

for the task at hand. As is typical with competitive tests, we cannot rule out the possibility that there exist models that yield better performances than the ones tested here. However, we have made an effort to include models that make extensive use of the available data, require high computational capacities, and take precautions that guard them against excessive error from variance. Therefore, these findings document the difficulty of predicting next-hour earnings, irrespective the available computational power.

As an additional test of these conclusions, we expand the forecasting window. So far, we have assumed that the task is to predict earnings one hour beyond trip end. However, it may rightly be pointed out that prediction over one hour may be difficult because it is too short a timescale. After all, even drivers who spend one hour taking a short fare and getting back in line waiting may get an airport trip the next time. Therefore, if the forecasting window is expanded and drivers form predictions over a longer time horizon, part of the randomness cancels out. We therefore repeat the analysis and expand the prediction window. Instead of predicting earnings one hour ahead, we use the same method to predict the average hourly earnings of the next two, three, four, or five hours. Although average next-hour earnings remain stable, the standard deviation is reduced, as are RMSE of most models. With a prediction horizon of five hours, the RMSE of the best predicting model improves 16 percent over the standard deviation. This result shows that predicting over longer time horizons allows the models tested here to yield better predictions. At the same time, these predictions remain too poor to conclude that next-hour earnings can be well predicted.

4.3 Conclusion: Prediction of Hourly Earnings is Difficult

This first analysis has demonstrated the difficulty for individual taxi drivers to predict their future earnings. The analysis has shown that the difficulty is not rooted in drivers' limitations in computational power but in a lack of predictability of their decision environment. This finding illustrates that rational expectations do not imply accurate predictions. Even if agents were to base their predictions on a state-of-the art machine learning model, the resulting predictions of their individual outcomes remain prohibitively imprecise.

The low level of predictability casts doubt on utility theory as a model of drivers' shift choices. Because utility maximization with imprecise expectations is unlikely to be rational, there remains little reason to expect shift choices to be consistent with utility theory. However, it remains true that a test of the assumptions cannot replace a test of the theory [Friedman, 1953]. However, as we have argued above, the existing evidence does not comprise a comparison of utility theory and alternative models in predicting taxi drivers' shift choices. The next section therefore addresses the question of how to best predict drivers' behavior.

5 How Are Drivers' Shift Ends Best Predicted?

At the heart of this text is the question of how drivers' shift ends are best predicted. Although Camerer et al. [1997] speculated about a target income, we have not found any test of this hypothesis. Instead, authors have devised and examined different versions of utility models that incorporate target incomes within its system of trade-offs. This narrow selection of models can answer the question of which version of utility theory is most consistent with observed data, but it does not yield conclusions about the usefulness of utility theory over alternative models. In the following, we present two of the utility models examined before and present four alternative models of shift termination..

The six models are then compared on their accuracy in out-of-sample prediction. Previous analyses have unanimously used in-sample fitting to draw conclusions about the descriptive power of different theories [e.g., Oettinger, 1999, Chou, 2002, Crawford and Meng, 2011, Farber, 2015]. However, the bias-variance trade-off implies that in-sample fit cannot be used to judge predictive accuracy, as these two are different measures. Following Friedman [1953], we argue that the ultimate purpose of theories is prediction of unobserved behavior and focus on an examination of their predictive accuracy. We therefore employ a competitive out-of-sample test that resembles the previous analysis. However, rather than focusing on the absolute performance of statistical models, we are now interested in the relative performance of the different behavioral models.

5.1 Candidate Strategies

We have selected the set of six models from different strands of literature and modeling approaches. The first approach, common in economics and much of statistics, models decision outcomes, typically at some level of aggregation. Models of this kind seek to approximate observed outcomes as well as possible, often using linear models with error terms. The second approach is common in cognitive science and machine learning and models the decision process. Models of this kind usually have no error terms and are deterministic in the sense that they output a specific decision rather than an average or an approximation[5].

We argue that process models are more natural models of human decisions than outcome models. These models take the information available to the agent as input and try to mirror the decision process. To this end, models need to be defined algorithmically, that is in the form of a decision strategy. In the case at hand, a decision strategy takes as input the information available to the driver at the end of each trip and arrives and decides whether the driver stops after this trip or not. The predicted shift end can then be compared to the observed shift end.

The existing literature on taxi drivers' shift ends proposes utility models, which are often implemented as outcome models [e.g., Farber, 2005]. These models specify the utility function, from which analysts can derive statistical models to link the probability of a trip end to environmental factors. The agents' decision process of comparing the utility of ending a shift to the expected utility of continuing the shift is assumed but not explicitly modeled. In particular, the process by which these expectations are formed is often left unspecified. To allow for a fair comparison among models, however,, the competitive framework of our analysis requires that models use the same input variables and output a prediction after each trip. Therefore, we had to implement the utility models

[5]Leo Breiman [2001] saw a similar distinction in modeling approaches in statistics, which he referred to as algorithmic and data modeling, respectively.

algorithmically and specify how agents form expectations. Among the many possibilities ranging from various forms of regression analysis to random forests, we devised an algorithm with no additional parameters that builds on the average of similar past shifts. In light of the fact that the PAST model in the previous analysis yielded almost identical performance as the REG model, we selected an algorithm that is conservative in the sense that the lack of additional parameters does not expose the strategy to additional variance. This way, we could devise an algorithmic version of all decision strategies considered here and presented in turn.

S1: Neoclassical Utility The first strategy considered here is the neoclassical utility model of intertemporal substitution, as presented by Crawford and Meng [2011]. According to this model, driver i compares utilities at the end of each trip t and terminates a shift as soon as the utility of terminating exceeds that of continuing. The utility of terminating the shift is calculated as follows,

$$U_t^T = r_t - \frac{\psi}{1+\upsilon} \times d_t^{1+\upsilon} \tag{8}$$

where r_t and d_t denote the shift earnings and duration at the end of trip t, respectively, and index i is suppressed for brevity. This termination utility is compared to the continuation utility, which gives the expected utility from not terminating the shift. The continuation utility is calculated as follows,

$$E_t[U^C] = E_t[r] - \frac{\psi}{1+\upsilon} \times E_t[d]^{1+\upsilon} \tag{9}$$

where $E_t[\cdot]$ denotes the expected value after trip t, r denotes shift earnings, and d denotes shift duration. To calculate the expected values of earnings and duration, driver i consults her recollection of comparable shifts. Among i's previous shifts of type m_t (day or night) with earnings at least equal to r_t and duration at least equal to d_t, similar ones are identified with the same values on demand shock, day of the week, and rain. That is, all previous shifts similar along these dimensions are identified and their shift earningss and durations are averaged to obtain the expected value for the shift at hand. In case, there is no previous shift with the same combination of demand proxies, individual variables are removed in reverse order until a recollection set of minimum size $N = 1$ is found. For example, if night shift s takes place on a rainy friday without demand shock but i has only experienced sunny fridays without demand shock, these fridays are used for comparison (rather than, say, rainy saturdays) because rain is the first variable to be ignored. In some rare cases, there is no comparison set because no previous shift of the same type was as long or as remunerative. In these cases, the expected values are calculated as $E_t[d] = d_t + 60$ and $E_t[r] = r_t + 60 \times \frac{r_t}{d_t}$, where d_t is measured in minutes. If $U_t^T > E_t[U^C]$, the shift is predicted to terminate after t, otherwise the procedure is repeated after the next trip, $t + 1$. If the predicted shift end is later than 24 hours after shift beginning, the prediction is truncated at 24 hours.

S2: Reference Utility The second strategy considered here follows the same process as the neoclassical utility strategy but uses a different utility function to compute U^T and $E_t[U^C]$. The utility function used here is described by

Crawford and Meng [2011] and based on Köszegi and Rabin [2006]. This function augments neoclassical utility, which they refer to as *consumption utility*, by a *gain/loss utility*, which depends on targets for both shift earnings and shift duration, as well as a parameter of loss aversion, λ, which reduces utility when earnings is below the earnings target or hours are above the duration target or both. In addition, a parameter η governs how relevant gain/loss utility is relative to consumption utility.

According to this model, the termination utility is calculated as

$$\begin{aligned} U_t^T = &(1-\eta) \times \left[r_t - \frac{\psi}{1+\upsilon} \times d_t^{1+\upsilon} \right] \\ &+ \eta \times \left[1_{(r_t - R \leq 0)} \times \lambda \times (r_t - R) + 1_{(r_t - R > 0)} \times (r_t - R) \right] \\ &- \eta \times \left[1_{(d_t - D \geq 0)} \times \lambda \times \left[\frac{\psi}{1+\upsilon} \times d_t^{1+\upsilon} - \frac{\psi}{1+\upsilon} \times D^{1+\upsilon} \right] \right] \\ &- \eta \times \left[1_{(d_t - D < 0)} \times \left[\frac{\psi}{1+\upsilon} \times d_t^{1+\upsilon} - \frac{\psi}{1+\upsilon} \times D^{1+\upsilon} \right] \right] \end{aligned} \tag{10}$$

where r_t and d_t denote shift earnings and duration at the end of trip t, respectively, R and D denote the earnings and duration targets, respectively, and index i is suppressed for brevity. Again, this termination utility is compared to the continuation utility, calculated as follows,

$$\begin{aligned} E_t[U^C] = &(1-\eta) \times \left[E_t[r] - \frac{\psi}{1+\upsilon} \times E_t[d]^{1+\upsilon} \right] \\ &+ \eta \times \left[1_{(E_t[r] - R \leq 0)} \times \lambda \times (E_t[r] - R) + 1_{(E_t[r] - R > 0)} \times (E_t[r] - R) \right] \\ &- \eta \times \left[1_{(E_t[d] - D \geq 0)} \times \lambda \times \left[\frac{\psi}{1+\upsilon} \times E_t[d]^{1+\upsilon} - \frac{\psi}{1+\upsilon} \times D^{1+\upsilon} \right] \right] \\ &- \eta \times \left[1_{(E_t[d] - D < 0)} \times \left[\frac{\psi}{1+\upsilon} \times E_t[d]^{1+\upsilon} - \frac{\psi}{1+\upsilon} \times D^{1+\upsilon} \right] \right] \end{aligned} \tag{11}$$

where $E_t[\cdot]$ denotes the expected value after trip t, r denotes shift earnings, and d denotes shift duration. To calculate the expected values during each shift s, driver i consults her recollection of comparable shifts.

Both versions of utility theory follow the conventional economic approach to decision modeling. Different attributes, here time and money, are brought onto a common scale, utility, and can be traded-off against one another. The exact trade-off is governed by a set of parameters that capture different preferences. In this theory, decisions respond to changes in demand if these changes are anticipated through the expected values of shift earnings and shift length.

S3: Earnings Target This third strategy is the first interpretation of a "raw" income target. In contrast to earlier decision models inspired by the earnings target, this model defines the target on the raw earnings scale, not on a utility scale. By implication, working hours and earnings are incommensurable and cannot be traded off against one another. The earnings target algorithm is defined as follows.

Driver i evaluates current shift earnings r_t at the end of each trip t. If $r_t > \rho_i$, the shift is ended and the end time of t is predicted to be the shift end.

Parameter ρ_i is individual to each driver and estimated from the data. If the predicted shift end is later than 24 hours after shift beginning, the prediction is truncated at 24 hours.

Strategies S1 – S3 were derived from the existing literature on taxi drivers' shift decisions. To expand the set of strategies tested, we generated additional satisficing strategies by varying the aspiration variable, that is the variable on which the aspiration level is defined. Whereas the earnings target terminates shifts as soon as cumulative shift earnings exceed their aspiration level, the following three heuristics use time on shift, clock hour, or the hiatus between trips to terminate a shift.

S4: Duration Target The fourth strategy considered here is the duration target. This strategy corresponds to a driver with a fixed shift duration planned at shift start, irrespective of information gathered and demand observed during the shift. The duration target algorithm is defined as follows.

Driver i evaluates current shift duration d_t at the end of each trip t. If $d_t > \delta_i$, the shift is ended and the end time of the previous trip is predicted to be the shift end. Parameter δ_i is individual to each driver and estimated from the data. If the predicted shift end is later than 24 hours after shift beginning, the prediction is truncated at 24 hours.

S5: Clock Target The fifth strategy considered here is the clock target. Similar to the duration target, the clock target strategy makes decisions based on time, but uses clock time rather than shift duration. When shifts consistently start at the same time, the two strategies lead to the same prediction. However, when shifts start at different times, ending after a specific shift length implies different clock times. The clock target algorithm is defined as follows.

Driver i evaluates current clock time c_t at the end of each trip t. If $c_t > \chi_i$, the shift is ended and the end time of the previous trip is predicted to be the shift end. Parameter χ_i is individual to each driver and estimated from the data. If the predicted shift end is later than 24 hours after shift beginning, the prediction is truncated at 24 hours.

S6: Hiatus Target The sixth and final strategy considered here is the hiatus target. Whereas the hiatus heuristic described in the previous section was defined on the hiatus between two purchases of the same customer, the hiatus target uses the hiatus between two subsequent trips, usually by different customers. The heuristic target algorithm is defined as follows.

Driver i evaluates the previous hiatus between two trips, h_t at the end of each trip t. If $h_t > \eta_i$, the shift is ended and the end time of the previous trip is predicted to be the shift end. Parameter η_i is individual to each driver and estimated from the data. If the predicted shift end is later than 24 hours after shift beginning, the prediction is truncated at 24 hours.

Strategies S3 to S6 are similar in structure but differ in their aspiration variables. The choice of the aspiration variable can be regarded as an expression of which goal the driver prioritizes. For example, an earnings target promises a fixed income, whereas a duration target promises a fixed shift length. Thus, a driver using a duration target may be interpreted as valuing a daily routine more than earning a fixed amount of income.

Such drivers may have family obligations that restrict them from extending their shifts beyond particular durations. With the exception of the hiatus target, the satisficing strategies essentially implement simple goals independently of demand. In contrast, it appears unrealistic that drivers have preferences over the hiatus between trips. Instead, they may use waiting time as a signal of demand: When demand decreases relative to the number of active taxis, waiting time should increase on average. The hiatus target therefore is the only satisficing strategy that responds to changes in demand.

To corroborate our selection of decision strategies, we conducted a survey among taxi drivers operating in the Hamburg taxi market. Drivers were recruited in two ways. First, taxi companies were approached by email and asked to forward the survey to their employees. Second, taxi drivers were approached in person in Hamburg on a sunny day in May outside of the observation window of the data. Participation in the survey was voluntary but incentivized with two prizes of €100, awarded at random. In total 70 drivers participated in the survey. Among other questions, they were asked to verbalize their strategies for calling it a day. For 27 drivers, the answers could not be classified because they were too unspecific (e.g., "too little demand") or because they referred to traffic and tiredness. Because traffic is either idiosyncratic or correlates with clock hour, these answers were ignored. The remaining answers were mapped onto the strategies above with 5 drivers mentioning target earnings, 6 drivers mentioning shift duration as a cue, 7 drivers mentioning clock hour as a cue, and 16 drivers mentioning the hiatus as a cue. None of the drivers verbalized the utility maximization strategy. In addition to verifying the existing candidate strategies, our second goal was to elicit strategies we have previously been unaware of. A total of 11 drivers mentioned new cues. These strategies included a target number of trips (1 driver), some form of minimum hourly wage (2 drivers), and some form of minimum number of trips per hour (2 drivers). Five drivers mentioned market saturation as a cue, mostly in the form of long lines of cars at taxi stands. Despite these somewhat scattered responses, it appears that there is no widely used strategy that our analysis ignores.

5.2 Empirical Approach

The purpose of this analysis is to identify for each driver the decision model that is most predictive of her observed shift ends. Our analysis differs from much of microeconomic analysis in several respects. We therefore present the empirical approach of this analysis in some detail and shift language from decision strategies to decision models to underline the descriptive question addressed in this section.

The empirical strategy in this section is an extension of the previous section's. As before, we use a competitive test to find the model yielding the best prediction for each shift. These predictions are then aggregated to classify drivers by their overall most predictive model. This approach does not require that the strategies are nested within the same class of models, such as linear models. Instead, models can be of different nature, provided they yield comparable outputs.

Algorithm 2 gives an overview of the testing procedure. To be able to classify drivers independently, the competitive test is carried out separately for each driver. As a first step, each shift is predicted by all six models and the best-predicting model is found for each shift. To this end, we employ 10-fold cross-validation. By this procedure, the sample of N shifts is divided into ten randomly composed folds of (roughly) equal size $N_f \approx \frac{N}{10}$ (step 1). All six models are then calibrated based on nine of these folds (step 2). For calibration, we use a derivative-free minimization algorithm [Hooke and Jeeves,

Algorithm 2
10-Fold Competitive Test of Models Predicting Shift Ends

```
1  forall shift type do
2      forall market do
3          foreach driver d do
4              1. assign shifts randomly across ten folds;
5              foreach fold f do
6                  2. calibrate all models based on all folds except f;
7                  foreach shift s in f do
8                      3. record end of final observed trip in s as actual end of s;
9                      foreach version v do
10                         foreach model m do
11                             foreach trips t do
12                                 4. use fitted models to calculate prediction;
13                                 if prediction is "continue" then
14                                     5. move to next t;
15                                 else
16                                     if end of t is later than 24 hours after begin of s then
17                                         6a. use begin of s + 24 hours as predicted end of s;
18                                     else
19                                         6b. record end of t as predicted end of s for model m;
20                                     end
21                                     7. move to next model;
22                                 end
23                             end
24                             8. calculate residual between predicted and actual end of s;
25                         end
26                     end
27                     9. calculate root mean squared residual, RMSR, across all 20 versions;
28                     10. identify best-predicting model by smallest RMSR;
29                 end
30             end
31             11. count for each model number of shifts it predicts best;
32             12. order models by count with m_1 predicting most shifts and m_6 fewest;
33             if count of m_1 > 1.2 × count of m_2 then
34                 13. classify d as using m_1;
35             else
36                 14. leave d unclassified;
37             end
38         end
39         15. count number of drivers for each model;
40     end
41 end
```

1961] that allows for a search of the best fitting set of parameters within given bounds. Whereas most parameters are left unconstrained, we constrain the search space for all utility parameters to be positive and the weight of gain-loss utility to $\eta < 1$. The calibrated models are used to calculate each model's predictions for the tenth fold (steps 3–7). For each model, the mean squared errors in these predictions are recorded (step 8). The procedure is performed ten times, each with a different fold used for prediction. Using this procedure, predictions can be calculated for all shifts without fitting.

Calculating the mean squared error for the sequential data is not straightforward. The predicted shift end can easily be computed when it lies within the observed shift. In contrast, it frequently occurs that a given model predicts that the driver continues beyond the observed shift end. After the shift ends, however, the models lack information to

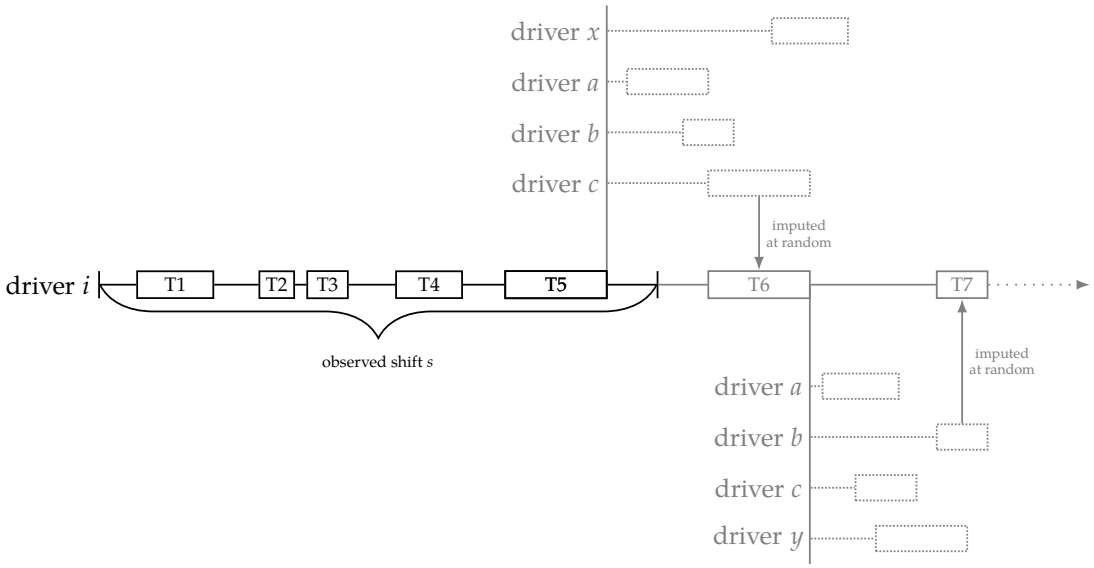

Figure 1: Construction of Extended Shifts: The shift of driver i lasts from trips T1 to T5; at the end of T5, drivers currently waiting for passengers are found and one of their next trips is chosen at random (here of driver c) and imputed as the next trip of i; at the end of that imputed trip, the procedure is repeated with drivers currently waiting; note that driver x is busy or off duty at T6 and not considered, whereas driver y is considered at T6 but was previously busy or not on duty.

be fed with, a problem typical of stopping decisions. Our solution exploits the size of the data sample. Although driver i ends shift s, we observe other drivers' trips in the aftermath of s. Although we do not observe the full market, the portion we do observe reflects variation in demand independently of i. In the absence of location data, we assume that each of the subsequent trips taken by competing drivers is equally likely to have been assigned to i, had she not ended her shift. Under these assumptions, we construct "extended shifts", that are shifts amended by imputed trips from other drivers.

Specifically, for each shift s we look at the final minute of the final trip and find all other drivers who fulfill three conditions: i) They are currently on duty, ii) they are currently without passenger, and iii) they complete at least one more trip during their current shift. Of these drivers, we select one at random and impute her next trip as the next trip of s, including trip begin, trip end, and trip earnings. This procedure is repeated for the final minute of this imputed trip until a total of fifty trips are imputed. The set of drivers from which the trips are imputed varies over time, as different drivers fulfill the above conditions at different points in time. The procedure is depicted in Figure 1. Using this procedure, we obtain hypothetical trip sequences beyond observed shift ends. Because each of these sequences consists of randomly imputed trips, the sequences beyond the observed section are random themselves. For this reason, we apply the selection procedure twenty times for each shift. The result is that for each shift s with n observed trips, we obtain twenty versions $s_1, s_2, ..., s_{20}$, each consisting of $n + 50$ trips, the first n of which are equal, whereas the following 50 trips vary. These extended shifts allow us to calculate predictions for all strategies beyond the observed shift end by applying each decision model to all twenty versions of each shift, and averaging the residuals in prediction (step 9). The model with the lowest root-mean squared residual

Table 4
Classifications of Shifts and Drivers

	All Drivers				Single Drivers only			
	Shifts		Drivers		Shifts		Drivers	
Model	Count	Share	Count	Share	Count	Share	Count	Share
neoclassical utility	77,452	9.4 %	7	0.2 %	26,194	9.6 %	4	0.5 %
reference utility	63,811	7.8 %	2	0.1 %	20,403	7.5 %	1	0.1 %
earnings	216,630	26.3 %	580	17.1 %	79,159	29 %	200	24.8 %
duration	254,593	30.9 %	1,482	43.6 %	76,942	28.2 %	228	28.3 %
clock	162,636	19.8 %	269	7.9 %	53,976	19.8 %	62	7.7 %
hiatus	48,364	5.9 %	5	0.2 %	16,568	6.1 %	1	0.1 %
unclassified	—	—	1,055	31 %	—	—	311	38.5 %
total	823,486	100 %	3,400	100 %	273,242	100 %	807	100 %

is then considered the best-predicting model for shift s (step 10).

The results are aggregated to classify drivers by their most predictive decision model. To this end, for each driver d, the number of shifts best predicted by each of the six models are counted. The first model, that is the model with the simple majority of shifts, is then selected as the best predicting model for d (steps 11 and 12), provided the simple majority is decisive. To be decisive, the first model needs to predict twenty percent more shifts than the second model. We take this precaution to ensure that drivers for whom two models predict roughly equally well remain unclassified (steps 13 and 14). Finally, classified drivers are counted (step 15).

5.3 Classification of Shifts and Drivers

We begin our presentation of results by the classification of shifts and drivers. Table 4 gives an overview of these classifications across all four samples we have analyzed separately. We start with an overview of shift classifications and then proceed with a discussion of driver classifications.

First, we focus on the classification of shifts. Columns 2 and 3 of Table 4 show the number of shifts best predicted by each of the six models, as well as their share. The majority of shifts are best predicted by the duration model, accounting for 30.92 percent of all shifts, as well as the earnings model, accounting for 26.31 percent of shifts, [6]. Because many drivers may need to adhere to a shift schedule, we also counted the number of shifts driven by single drivers who are likely unconstrained by shift schedules. Columns 6 and 7 show that the duration model accounts for a smaller percentage of shifts but also the majority of single driver shifts is best predicted by the duration and earnings models.

Of the remaining shifts, most were predicted by the clock model and fewest by the hiatus model. Again, this result holds for all drivers and the subgroup of single drivers. Similarly to the hiatus model, both utility models account for at most 10 percent of all shifts. Between them, the neoclassical utility model predicts somewhat more shifts than the reference utility model. In contrast to the hiatus and utility models, the clock model typically accounts for about 20 percent of shifts. Overall, it appears that a distinction can be made between the duration, earnings, and clock models jointly predicting around

[6]Note that shift classifications are not independent of drivers: By itself, each shift end is consistent with all six models if model parameters could be chosen freely. However, if each model's parameter values are fixed for each driver, we can identify for each shift the model yielding the best prediction. The counts of these models are presented here.

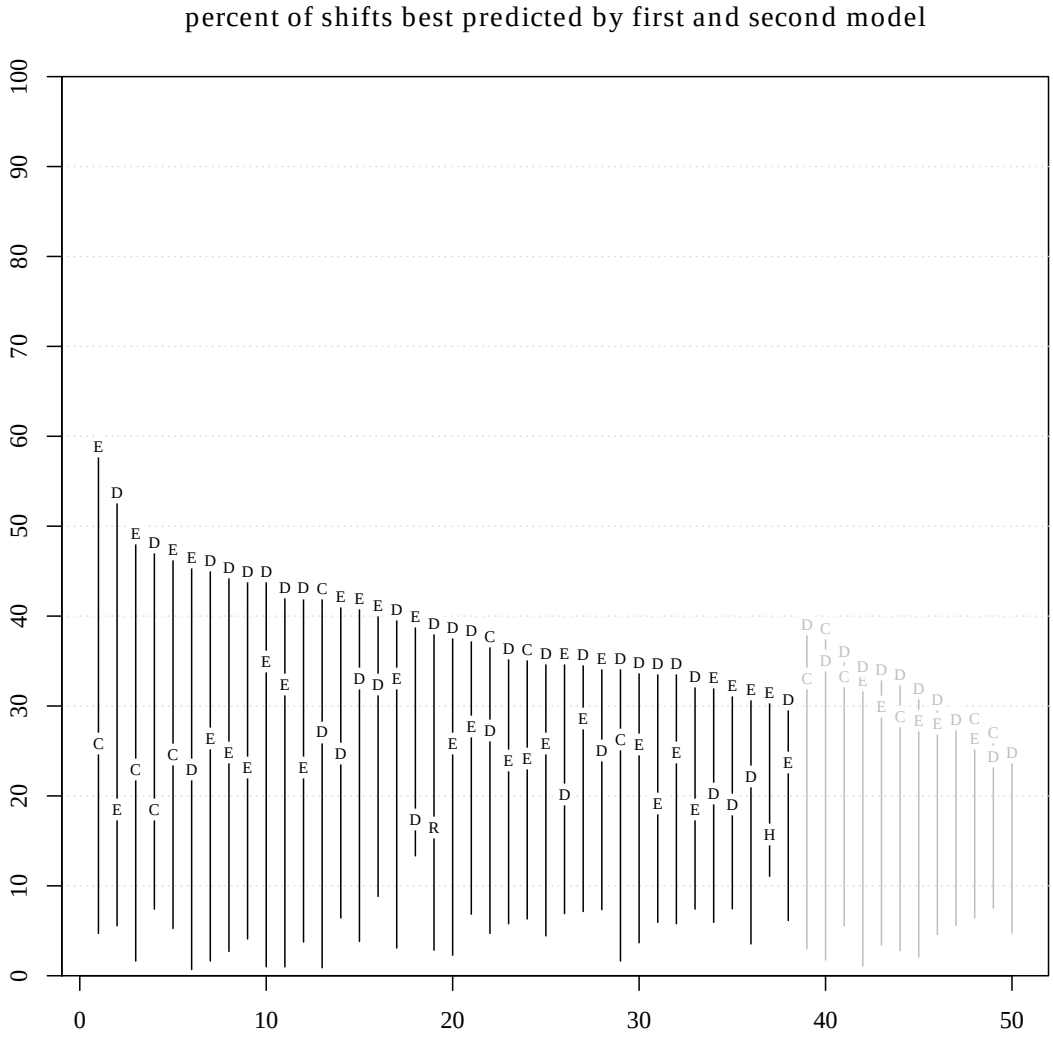

Figure 2: Range in percentages of shifts predicted by six models for random sample of 50 drivers; letters indicate model with highest and second highest number of shifts: neoclassical utility (U), reference utility (R), earnings (E), duration (D), clock (C), hiatus (H); drivers classified by their first model shown in black. unclassified drivers shown in grey.

three quarters of shifts and the hiatus and utility models predicting around one quarter of shifts.

Next we examine the classification of drivers. To arrive at these classifications, we calculated for each driver the percentage of shifts best predicted by each of the six models. The vertical bars in Figure 2 show for a random sample of 50 drivers the range of these percentages. The letters on top indicate the models predicting the highest number of shifts, which we refer to as the first and second model, respectively. The lower end of the line shows the percentage best predicted by the last model. Differences between these two models are stark and mostly above 25 percentage points.

Drivers were only classified by their first model if it predicted 20 percent — not percentage points — more shifts than the second model, otherwise they were left unclassified. In Figure 2, the first 38 drivers were classified according to their first model, whereas the latter 12 drivers were left unclassified. For these drivers, we deemed the evidence too weak for a classification. In Table 4, columns 4 and 5 of Table 4 show the number and share of drivers classified by each model. Around 30 percent of all drivers

Table 5
Wage Elasticity of Labor Supply Across All Shifts

	All Drivers			Single Drivers		
Variable	Estimate	Standard Error	p-Value	Estimate	Standard Error	p-Value
Day Shifts						
intercept	2.885	0.017	<0.001	2.640	0.031	<0.001
log wage	−0.206	0.006	<0.001	−0.089	0.012	<0.001
rainy day	−0.000	0.001	0.716	−0.005	0.002	0.036
Saturday	−0.069	0.002	<0.001	−0.113	0.004	<0.001
Sunday	−0.058	0.002	<0.001	−0.074	0.004	<0.001
shock	−0.012	0.002	<0.001	−0.011	0.003	<0.001
Night Shifts						
intercept	2.206	0.016	<0.001	2.132	0.039	<0.001
log wage	−0.002	0.006	0.692	0.051	0.015	<0.001
rainy day	−0.010	0.002	<0.001	−0.013	0.004	0.004
Saturday	−0.030	0.003	<0.001	−0.159	0.007	<0.001
Sunday	0.094	0.003	<0.001	0.034	0.007	<0.001
shock	−0.005	0.002	0.034	−0.014	0.005	0.009

were left unclassified, with a slightly higher percentage among single drivers.

Among those drivers we could classify, consider first the earnings and duration models. Again, these two models jointly account for the majority of drivers, including those left unclassified. By these results, the duration model offers the widest description of drivers' shift choices. Consider next the clock model, which accounts for about 10 percent of all drivers. Together with the duration and earnings models, the clock model accounts for around two thirds of drivers. With around 30 percent unclassified, the hiatus model and the utility models each account for a fraction of a percent of drivers. These results hold true in our subsample of single drivers, although the duration model accounts for a smaller percentage of drivers, indicating that part of its descriptive may be due to shift schedules. Overall, driver classifications therefore mirror the dichotomy of shift classifications, with stronger divergence in the performances of both groups of models.

Table 4 shows results aggregated across all drivers, ignoring differences between day and night shifts and between the old and the new market. In Appendix C we produce a more extended version of Table 4 that reports results separately for the four samples constructed by crossing shift type and market. Although there are some differences, primarily between day and night drivers, the results do not change qualitatively. In addition, we report more details on the number of shifts consistent with driver classifications, on average between 30 and 40 percent per driver, indicating that drivers may be best predicted by using a combination of models. We therefore examine the second model and find that for most drivers, the majority of shifts is best predicted by the duration and earnings targets but percentages vary. In addition, Appendix B reports histograms of parameter estimates.

5.4 Aggregate Outcomes

Given the classifications of shifts and drivers above, we can compare drivers' aggregate outcomes stratified by their best predicting model. We begin with a short discussion of the labor supply elasticity and then turn to average hourly earnings.

To obtain the overall wage elasticity, we have followed the IV-approach by Farber

Table 6
Elasticities And Hourly Earnings by Model

	Across Drivers				Across Shifts			
			Earnings				Earnings	
Model	Drivers	Elasticity	Mean	SD	Shifts	Elasticity	Mean	SD
neoclassical utility	7	0.99	21.4	7.9	77,452	0.00	18.7	7.9
reference utility	2	−0.52	19.8	6.8	63,811	−0.24	19.1	7.4
earnings target	580	−0.76	15.5	5.5	216,630	−0.86	17.3	7.1
duration target	1,482	0.05	19.4	5.1	254,593	−0.00	18.5	7.9
clock target	269	0.09	20.2	5.1	162,636	0.08	18.3	7.4
hiatus target	5	−0.06	22.8	8.0	48,364	−0.01	16.4	9.5

Notes: Across drivers: hourly earnings averaged across consistent shifts and drivers, hourly earnings SD gives SD across drivers, elasticity gives median elasticity estimate from IV regression for each driver; across shifts: hourly earnings averaged across shifts of the same strategy, elasticity obtained from regression using all shifts of the same strategy; for brevity, no separate display for single drivers.

[2015] and regressed `ln(sdur)`, the natural logarithm of shift earnings, on `ln(searn/sdur)`, `rain.d`, `saturday`, and `sunday` separately for day and night shifts. In addition, we have added the variable `shock`, a dummy indicating one of the demand shocks listed in Table 1. We used other drivers' average hourly earnings on the same day as an instrument for `ln(searn/sdur)`. This approach is similar to Farber's approach, who has used a non-overlapping sample to instrument wages of drivers in the remainder of his data set.

The results are shown in Table 5 with elasticity estimates of −0.206 and −0.002 for day and night shifts, respectively. For day drivers, we therefore find a negative wage elasticity, despite using an instrumental variable. For night drivers, we find a wage elasticity very close to zero, implying that on average, drivers do not respond to wage increases. The elasticities of single drivers are negative for day shifts and positive for night shifts but small in magnitude.

With heterogeneous drivers, calculating elasticities across all drivers may be misleading and our classification of drivers allows for a more detailed examination. The estimates in Table 5 give an aggregate summary across all drivers but also hide a considerable amount of variation between them. To shed light on the relation between best predicting model and demand elasticity, we can use both the classifications of shifts and drivers. The former give a complete picture, whereas the latter give a clearer picture of the relation between shift termination model and aggregate outcomes. Table 6 reports both of these analyses.

First, we estimated the elasticity separately for each driver by applying the IV-approach above those shifts that are consistent with the driver's classification. Column 3 of Table 6 reports these results, stratified by model. Consistently with expectations, the median elasticity of drivers best predicted by the earnings target is negative. In contrast, the median elasticity of drivers best predicted by the duration and clock targets fluctuate around zero and are small in magnitude. This is consistent with expectations as shifts in these models are terminated independently of their profitability. Because drivers best predicted by the utility and hiatus models are only few in number, we refrain from interpretations of their median elasticities.

In a second analysis, we have applied the IV analysis to all shifts, irrespective of driver classifications and the results are shown in column 7 of Table 5. Again, we find a negative elasticity for shifts best predicted by the earnings model, whereas those best predicted by the duration and clock models are close to zero on the median. The neoclassical utility

model yields an average elasticity parameter of zero but averages for both utility models models differ considerably from the analysis across drivers. We therefore suspect that these results reflect primarily between-driver variability as these models typically predict only few shifts per driver. For brevity, both of these analyses have ignored difference between day and night shifts, as well as between the old and the new market. Appendix D reports aggregate outcomes separately for these four subsamples.

One final issue concerns the profitability of different models, as measured by the hourly earnings they generated. Mean hourly earnings across drivers are given in columns 4 and 5 of Table 6. Drivers best predicted by the earnings model exhibit the lowest mean earnings, whereas those best predicted by the hiatus model exhibit the highest. Columns 8 and 9 of Table 6 report means across shifts rather than drivers. By comparison, differences between models are somewhat smaller but the earnings model yields, on average, lower hourly earnings than the duration model. Similarly, the heuristic models appear to yield somewhat lower mean earnings than the utility models. At the same time, these differences are small compared to the variation within each class. For a more detailed illustration, Appendix D reports kernel density plots of mean hourly earnings, separately for day and night shifts in the old and new market.

5.5 Conclusion: Duration and Earnings Models Most Predictive

The second analysis has demonstrated the inability of utility theory to predict the majority of shifts for a meaningful number of drivers. Although both utility models could predict sizable portions of shift ends, these shifts are spread across many drivers rather than concentrated on drivers that can be described consistently by these models. Instead, those models of drivers as pursuing simple aspirations on earnings or time are best at predicting drivers, despite the fact that these models assumed drivers to have constant aspiration levels within each sample of shifts, irrespective of predictable demand shocks. In particular, many drivers are best predicted by some combination of earnings and duration targets. We hypothesize that the individual mix depends on personal circumstances. Across models, wage elasticities were, with few exceptions, found to be as expected. The distributions of average hourly earnings are very similar across models with heuristic models exhibiting slightly lower modes than utility models in some samples.

6 Discussion

This paper set out to predict taxi drivers' earnings and shift ends. For both analyses, we have used competitive out-of-sample tests. As we have argued, this methodology differs from conventional microeconomic analysis but has desirable properties for empirical work. It directly tests models' predictive powers rather than their abilities to adjust to existing data. Further, the competitive approach enables a comparison of different, non-nested models, provided they work on the same input and yield comparable outputs. This methodology has resulted in a picture of taxi drivers' choices that differs from earlier results in some important points.

Our first analysis has found that a variety of statistical models can hardly predict individual drivers' hourly earnings from readily observable variables. Unless we are willing to assume that drivers' predictive talents exceed the power of these models, we may accept the conclusion that drivers' earnings of the next hour appear random. These findings contradict those by Farber [2015] and lead to the divergent conclusion that even in reference utility models that set their reference point at the expected level of earnings,

there is ample room for the reference point to affect decisions. However, the level of predictability is so low that behavioral models relying on expected earnings, such as neoclassical and reference-dependent utility models, seem unfit for the task at hand. This conclusion is conditioned on the assumption that our set of considered covariates is exhaustive and drivers do not have and use information beyond these variables. In practice, of course, there can exist additional variables that drivers could use for making predictions. Such variables include information on the timing of specific events, such as concerts or sports events, or the number of other taxis on shift. This being said, we also note that the small progress made with the existing set of variables makes us skeptical that extensions will lead to improvements so large as to qualitatively change our conclusion that earnings predictions are difficult in practice and prone to error. Because the merit of utility theory is typically seen in its assumed rationality, our findings leave no theoretical reasons to assume a priori that utility theory describes drivers' behavior better than any other theory.

Our second analysis found that most drivers are best described by one of three satisficing models that set aspiration levels on earnings, shift duration, or clock time. These aspiration levels are fixed in the sense that drivers may use different aspiration levels for day and night shifts but ignore other factors. Camerer and colleagues [1997] had referred to such models as the "strong form of the target income hypothesis", emphasizing their inflexibility. However, given the small predictive power of observables such as weather or demand shocks, we decided to test models with fixed aspiration levels. Noticeably, these inflexible models yielded considerably better predictions than both utility models we have tested.

Importantly, we cannot conclude from these results that drivers actually use one model rather than another. Instead, there are two alternatives. First, it is possible that they use a model not tested that would yield better results than the models included here. It is important to note that results of our analysis are relative to the selection of candidate models. Although we have made an effort to select promising and relevant candidates, we cannot exclude that there are better predicting models from the universe of countless possible models. Second, it is theoretically possible that drivers use one of the utility models but are better predicted by the satisficing models for their lower exposure to variance. For illustration, Artinger et al. [under review] describe a scenario with high variance in which a regression model with two variables yields better predictions of data generated using three variables than the data-generating model itself. Despite these caveats, our analysis has demonstrated a considerable predictive advantage of satisficing models over utility models.

The results of this study hold true for the Hamburg taxi market and generalizations must be made with caution. Markets differ along several dimensions, including demand patterns as well as drivers' goals and working conditions. The findings indicate that drivers use a set of heuristics and one can assume that these strategies are not selected at random but according to the specifics of each decision environment. For example, Farber [2015] reports positive wage elasticities for the majority of NYC taxi drivers, indicating that conditions differ between New York and Hamburg, leading drivers to use different strategies[7]. Therefore, similar analyses for other markets, for taxis or otherwise, seem more sensible than over-generalizations beyond the domain studied here.

Our findings also imply that positive elasticities are more difficult to attain than is

[7]Similarly, Fehr and Goette [2007] report positive wage elasticities for bike messengers in their experiment. Unlike Hamburg taxi drivers who spend most of their shifts waiting for passengers, the authors stress that bike messengers have considerable control over their earnings the effort they exert.

commonly understood. The inaccuracy of predicted earnings virtually prohibits any judgements of future earnings as worth the time or not. To illustrate, consider a taxi driver on a small island who operates one of few taxis that bring day tourists from the harbor into town and back. The number of tourists varies with a few observables such as weather, day of the week and time of the year. Because day tourists arrive by ferry, the driver can use the ferry schedule to predict the timing of passengers across the day and terminate the shift when expected earnings become too low. Therefore, the driver can focus work on the most profitable hours and attain a positive elasticity. The conditions for Hamburg taxi drivers are quite different. Trips do not necessarily go from harbor to town and vice versa but vary in length and profitability, and timing of passengers cannot be looked up on a schedule. Individual earnings therefore not only depend on overall demand but vary along many factors including location and the labor supply of other drivers. Predicting market averages therefore helps little in finding the best point during the day to terminate the shift. Although such points may exist, drivers may find themselves unable to identify them.

Nonetheless, positive elasticities are entirely possible. For example, a driver may decide a priori to work three hours a night during the week and six hours on profitable weekend nights. Presumably, this simple strategy generates a positive wage elasticity and illustrates how the wage elasticity is also affected by shift choice and shift beginnings. Indeed, other strategies are conceivable that can generate positive elasticities, such as the hiatus heuristic that seeks to detect signals of decreasing demand. However, our findings suggest that substantially positive elasticities are difficult to attain based on prediction of future earnings.

The emergent picture of the taxi market is more complicated that initially assumed by Camerer and colleagues. The authors had chosen the taxi markets for its variability in daily earnings, which could be used to estimate the wage elasticity. This plan ignored the fact that the variability in earnings appears to be dominated by factors not readily observable, which makes positive elasticities difficult to attain based on prediction of earnings. Under these circumstances, it appears misguided to assume a positive wage elasticity to be rational and view a negative wage elasticity as evidence against neoclassical theory.

We also note that the focus on elasticities can be inadequate for other reasons. Consider again the driver working three hours on weeknights and six on weekend nights. If the same driver were to stop working on weeknights altogether, the elasticity likely decreases as there are fewer comparatively short, comparatively unprofitable shifts. At the same time, average wages would likely increase. This discrepancy illustrates that evaluating drivers based on a single indicator can be misleading and does not do justice to the intricacies of preferential choice. Given the lack of evidence that violations of coherence, the classical criterion for economic rationality, impair choices [Arkes et al., 2016], we advocate a more comprehensive benchmark of ecological rationality that judges decision strategies by their ability to reach defined goals. By this account, an analysis of the rationality of drivers' strategies requires detailed data on drivers' goals. Until such data is available for analysis, assertions of rationality remain speculative.

Finally, our analysis has shown the limits of conventional decision modeling. One of the reasons the income-target hypothesis has been criticized is the counter-intuitive idea that drivers would *prefer* a specific income. On the one hand, why would income beyond a target be less valuable than below the target? On the other hand, the conventional approach models deviations from neoclassical theory through modifications of the preference structure. In this article, we have described an alternative explanation. By

this hypothesis, drivers have no particular preference for any given amount of earnings. However, the uncertainty of their decision environment prevents them from calculating optimal paths of action. Therefore, they rely on a toolbox of satisficing strategies. The selection of the aspiration variable can be regarded as an expression of which goal the driver finds most important. However, the aspiration level is not necessarily an expression of preference but the value that yields the best trade-off of time and earnings in the driver's specific decision environment. The empirical challenge lies in an understanding of the circumstances under which such strategies are rational.

References

Hal R Arkes, Gerd Gigerenzer, and Ralph Hertwig. How bad is incoherence? *Decision*, 3 (1):20, 2016.

Florian M. Artinger, Gerd Gigerenzer, and Perke Jacobs. Sixty-five years of satisficing: Integrating two traditions. under review.

Leo Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231, 2001.

Colin Camerer, Linda Babcock, George Loewenstein, and Richard Thaler. Labor supply of new york city cabdrivers: One day at a time. *The Quarterly Journal of Economics*, 112 (2):407–441, 1997.

Yuan K Chou. Testing alternative models of labour supply: Evidence from taxi drivers in singapore. *The Singapore Economic Review*, 47(01):17–47, 2002.

Vincent P Crawford and Juanjuan Meng. New york city cab drivers' labor supply revisited: Reference-dependent preferences with rational-expectations targets for hours and income. *American Economic Review*, 101(5):1912–32, 2011.

DWD Climate Data Center. Historical daily station observations (temperature, pressure, precipitation, sunshine duration, etc.) for Germany, 2018a. version v006.

DWD Climate Data Center. Historical hourly station observations of precipitation for Germany, 2018b. version v006.

Henry S Farber. Is tomorrow another day? the labor supply of new york city cabdrivers. *Journal of Political Economy*, 113(1):46–82, 2005.

Henry S Farber. Reference-dependent preferences and labor supply: The case of new york city taxi drivers. *American Economic Review*, 98(3):1069–82, 2008.

Henry S Farber. Why you can't find a taxi in the rain and other labor supply lessons from cab drivers. *The Quarterly Journal of Economics*, 130(4):1975–2026, 2015.

Ernst Fehr and Lorenz Goette. Do workers work more if wages are high? evidence from a randomized field experiment. *American Economic Review*, 97(1):298–317, 2007.

Milton Friedman. *Essays in Positive Economics*. The University Press of Chicago, Chicago, 1953. ISBN 0226264033.

Stuart Geman, Elie Bienenstock, and Rene Doursat. Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, 4:1–58, 1992. doi: 10.1162/neco.1992.4.1.1.

Gerd Gigerenzer and Daniel G. Goldstein. Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review*, 103(4):650–669, 1996. ISSN 0033-295X.

Patrick Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, New York, NY, 2nd edition, 2001.

Robert Hooke and Terry A Jeeves. A "Direct Search" solution of numerical and statistical problems. *Journal of the ACM (JACM)*, 8(2):212–229, 1961.

Frank H. Knight. *Risk, Uncertainty, and Profit*. Houghton Mifflin, Boston, 1921.

Botond Köszegi and Matthew Rabin. A model of reference-dependent preferences. *The Quarterly Journal of Economics*, 121(4):1133–1165, 2006.

Sarah Levy. Bitte aussteigen [please exit here]. *Die Zeit*, 2015(13), 2015.

Kurt Lewin, Tamara Dembo, Leon Festinger, and Pauline Snedden Sears. Level of Aspiration. In Joseph McVicker Hunt, editor, *Personality and the Behavior Disorders*, pages 333–378. The Ronald Press Company, New York, 1944.

Rodney Maddock and Michael Carter. A child's guide to rational expectations. *Journal of Economic Literature*, 20(1):39–51, 1982.

John F Muth. Rational expectations and the theory of price movements. *Econometrica: Journal of the Econometric Society*, pages 315–335, 1961.

Gerald S Oettinger. An empirical analysis of the daily labor supply of stadium venors. *Journal of political Economy*, 107(2):360–392, 1999.

Judea Pearl. *Intelligent search strategies for computer problem solving*. Addision Wesley, 1984.

Paul A Samuelson and William D Nordhaus. *Economics*. McGraw-Hill, 16 edition, 1998.

David C Schmittlein, Donald G Morrison, and Richard Colombo. Counting your customers: Who-are they and what will they do next? *Management science*, 33(1):1–24, 1987.

Herbert A Simon. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1):99–118, 1955.

Herbert A Simon. Rational choice and the structure of the environment. *Psychological Review*, 63(2):129, 1956.

Herbert A. Simon. Rational Decision Making in Business Organizations. *The American Economic Review*, 69(4):493–513, 1979.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131, 1974.

Markus Wübben and Florian v Wangenheim. Instant customer base analysis: Managerial heuristics often "get it right". *Journal of Marketing*, 72(3):82–93, 2008.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

Appendix A Plots of Shift Begins

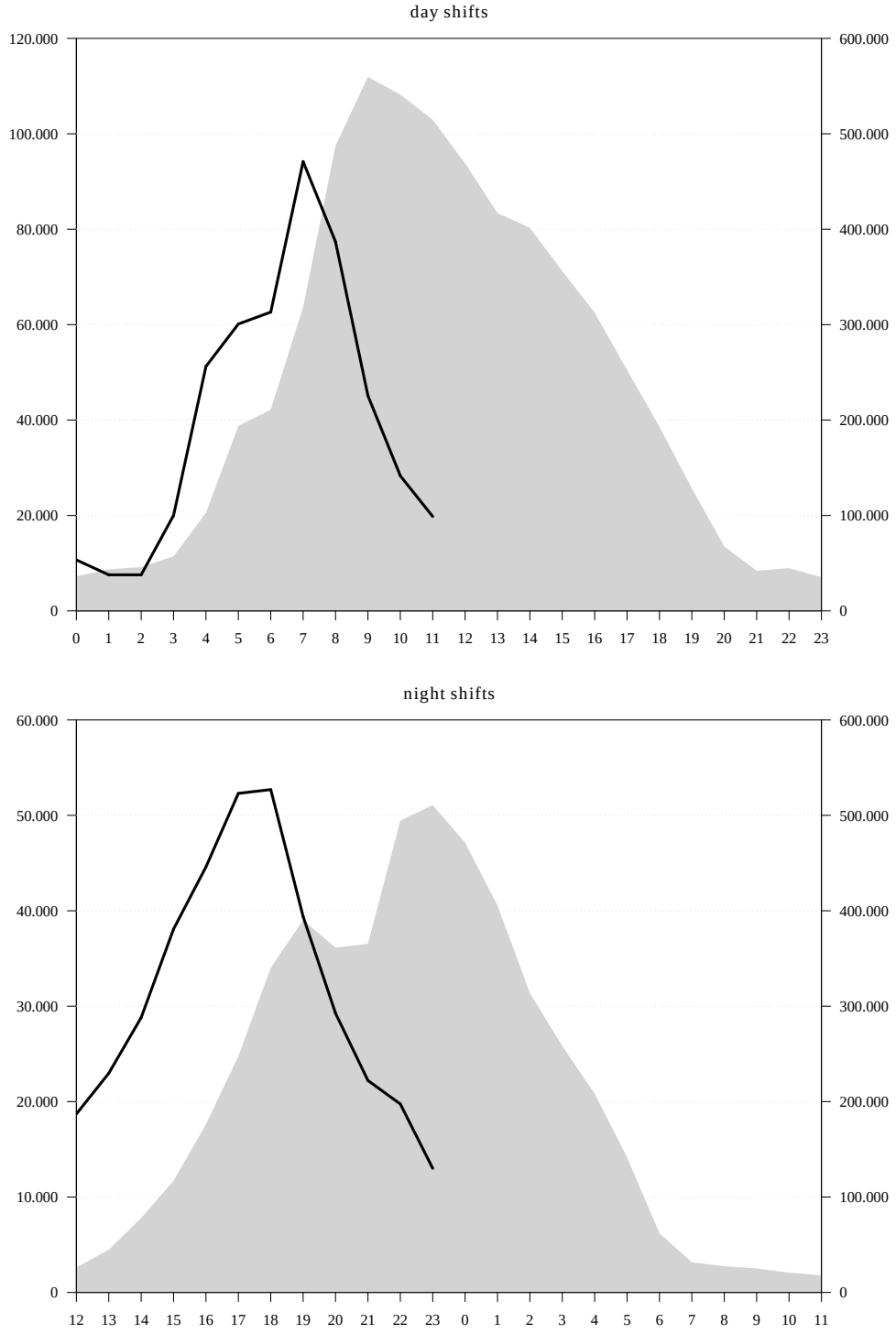

Figure A1: Number of shift begins (black, left axis) and trips (grey, right axis) across clock hours.

Appendix B Parameter Estimates

The classifications reported in Section 5 rest crucially on the plausible parametrization of the six models. We therefore include a brief description of the estimated parameter values. Although each model was evaluated for each driver, we examine only drivers we could classify and for each of those, we examine only shifts that are consistent with the driver's classification. These parameters are estimated ten times per driver, once for each of the ten folds across which the driver's shifts were distributed. We have computed for each driver the median parameter value across these ten folds. The distributions of these median values are shown in Figures B1 to B4, respectively. We display them separately for day and night shifts and for the old and new markets, that is before and after the fare increase and the introduction of the minimum wage.

In our implementation, neoclassical utility theory has two free parameters that were estimated from the data. First, the disutility of work was restricted in fitting to $0 \leq \theta$ and was estimated for the 7 day drivers in the new market at values between 0.437 and 1.313 and for the 7 drivers in the old market at values between 0.496 and 3.124. Values for night drivers ranged from 0.059 to 1.630 for both markets. Second, the wage elasticity parameter was restricted in fitting to $0 \leq \rho$ and was estimated for most drivers across markets and shift types above one. However, because there were few drivers best described by the neoclassical utility model, these results cannot be used to infer representative parameter estimates. Similarly, there were fewer than five drivers best described by the reference utility model across all four samples and we refer the reader to Figures B1 to B4 for distributions of its parameters.

We now turn to the four satisficing models, each of which has one parameter only. For day drivers best predicted by the earnings model, the parameter estimates follow a bell-shaped distribution, irrespective of old or new market. The central 60 percent of drivers between the 20^{th} and 80^{th} percentiles had estimated targets between €111 and €187. For night drivers, the distribution was flatter with 60 percent of drivers estimated to have targets between €89 and €196. Parameter estimates for drivers best predicted by the duration model followed a bell-shaped distribution with 60 percent of estimates between 447 and 605 minutes for day drivers and between 408 and 561 minutes for night drivers, irrespective of the market. For day drivers best predicted by the clock model, 60 percent of targets were estimated between 1pm and 8pm, with no apparent differences between old and new market. In contrast, 60 percent of night drivers had targets estimated between 9pm and 6am, again with no apparent differences between the two markets. For those day drivers best predicted by the hiatus model, 60 percent of drivers had estimated targets between 44 and 61 minutes. Similarly, 60 percent of night drivers had estimated targets between 36 and 54 minutes.

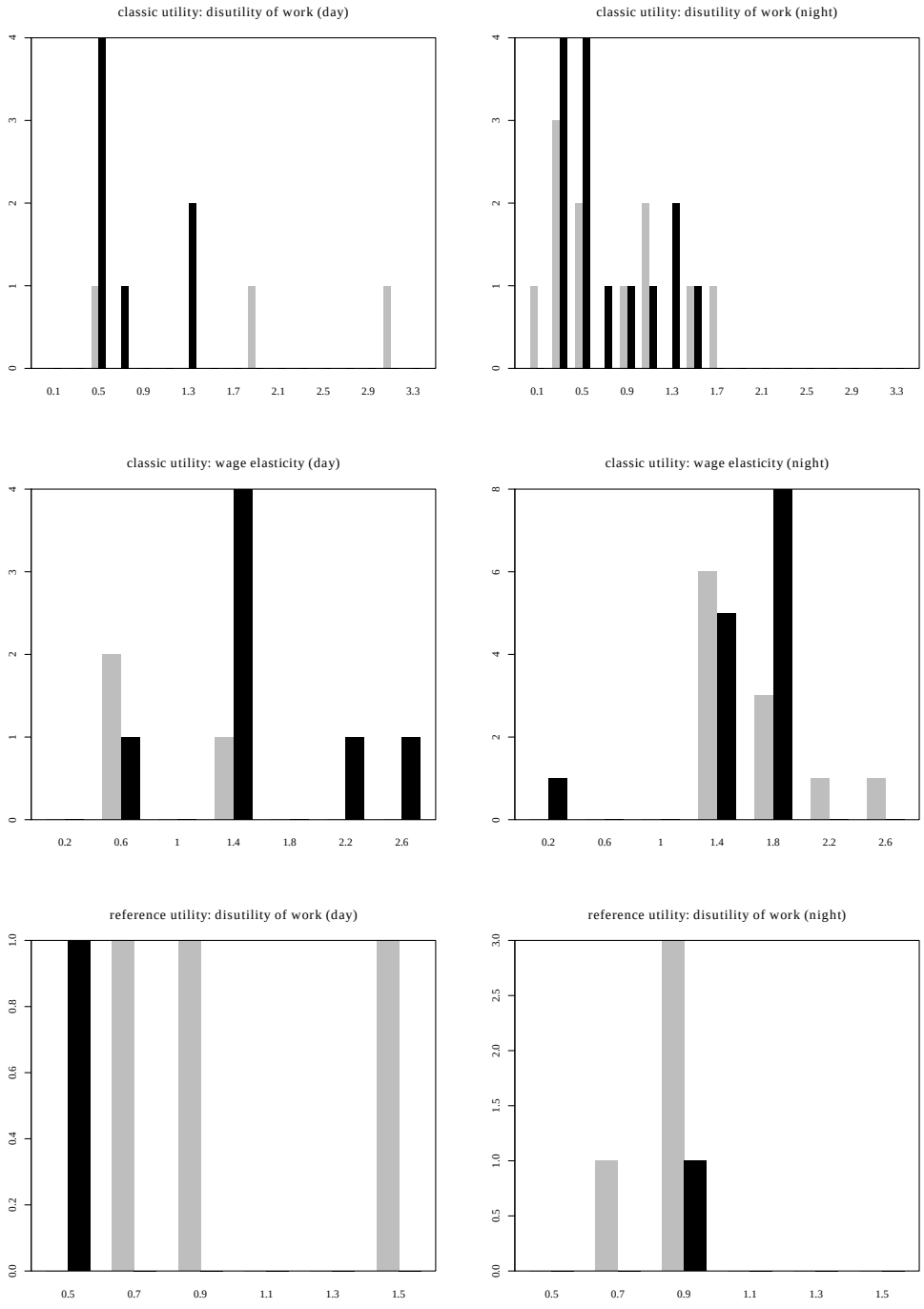

Figure B1: Histograms of median parameter estimates for old market (grey) and new market (black).

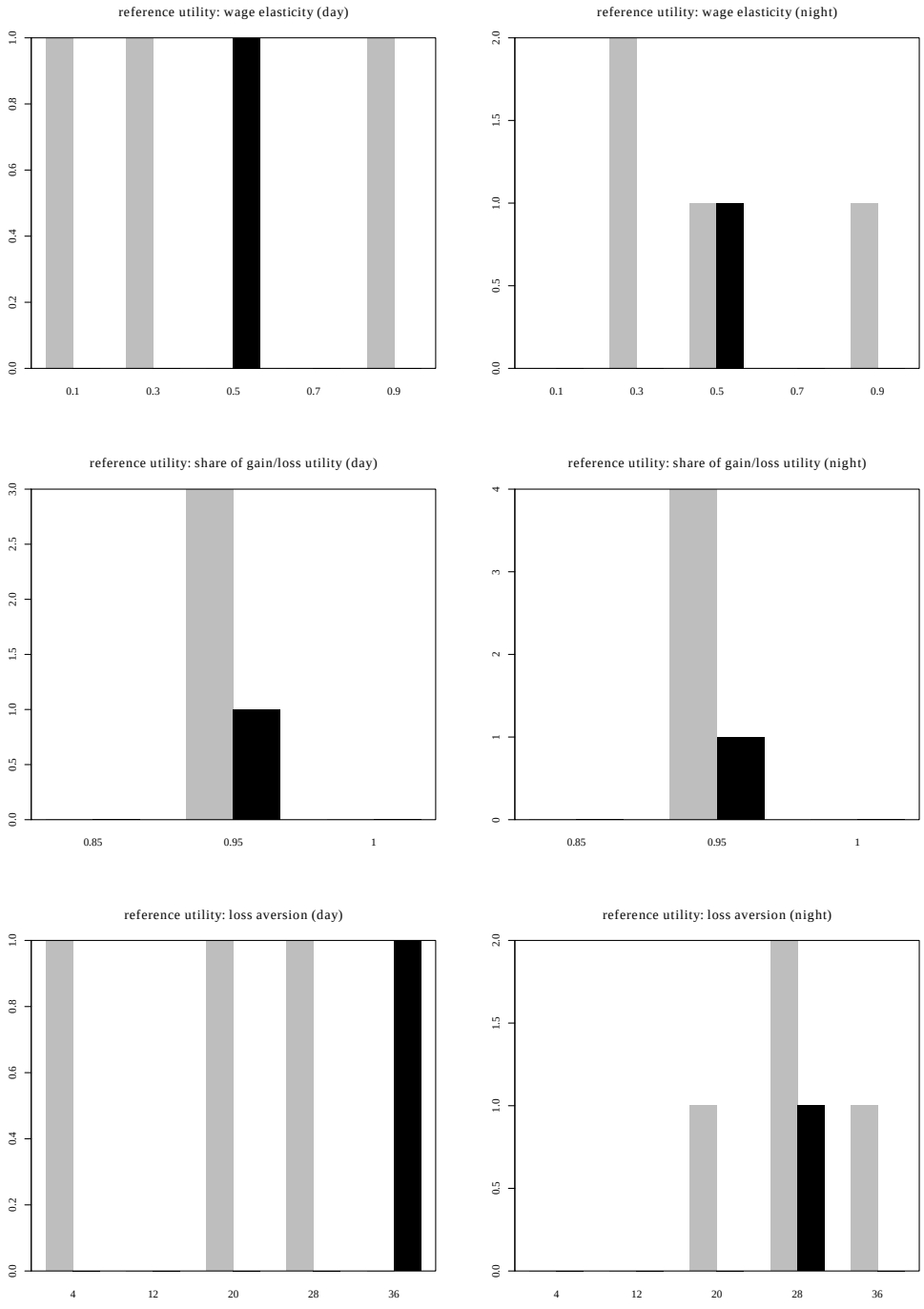

Figure B2: Histograms of median parameter estimates for old market (grey) and new market (black).

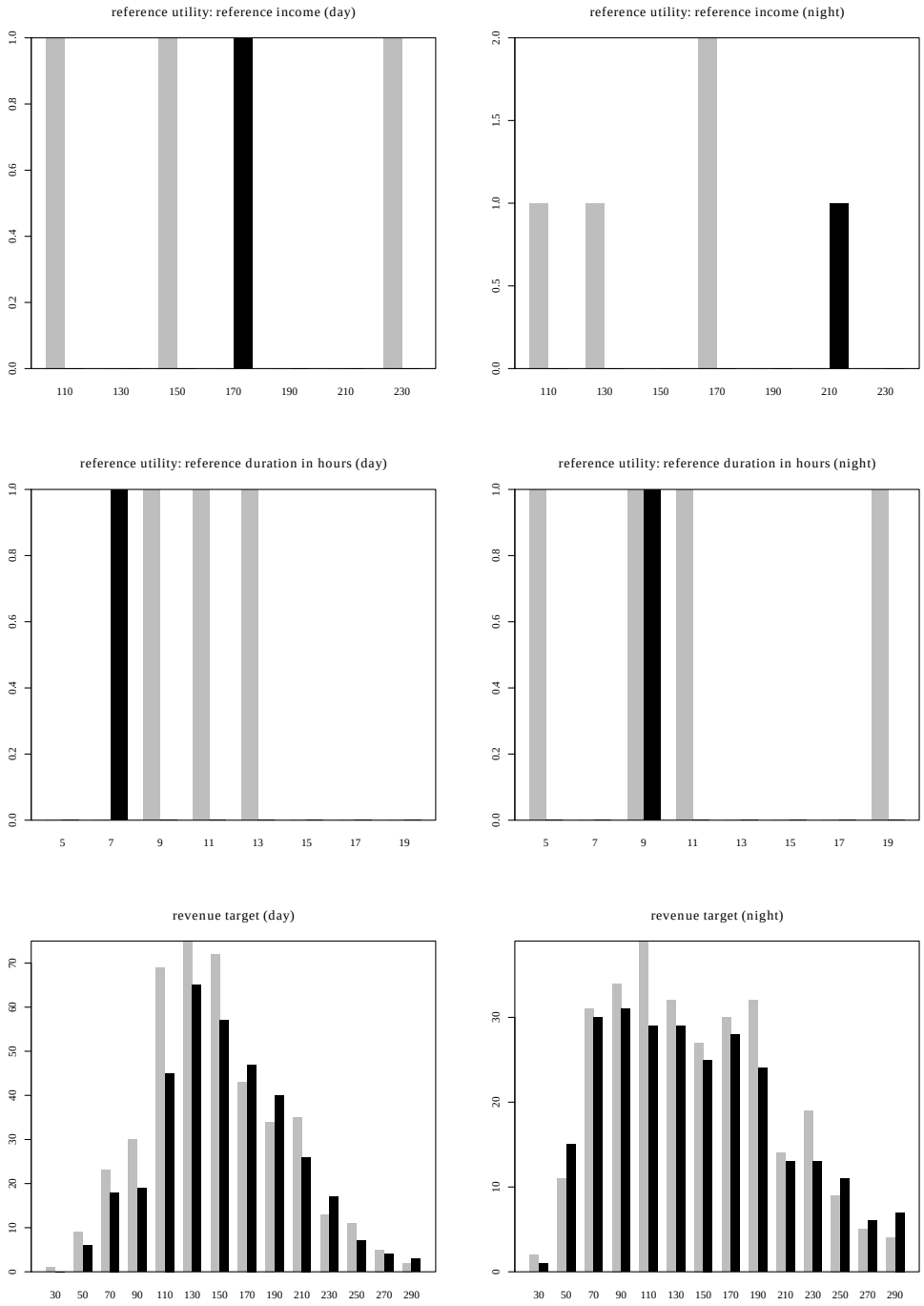

Figure B3: Histograms of median parameter estimates for old market (grey) and new market (black).

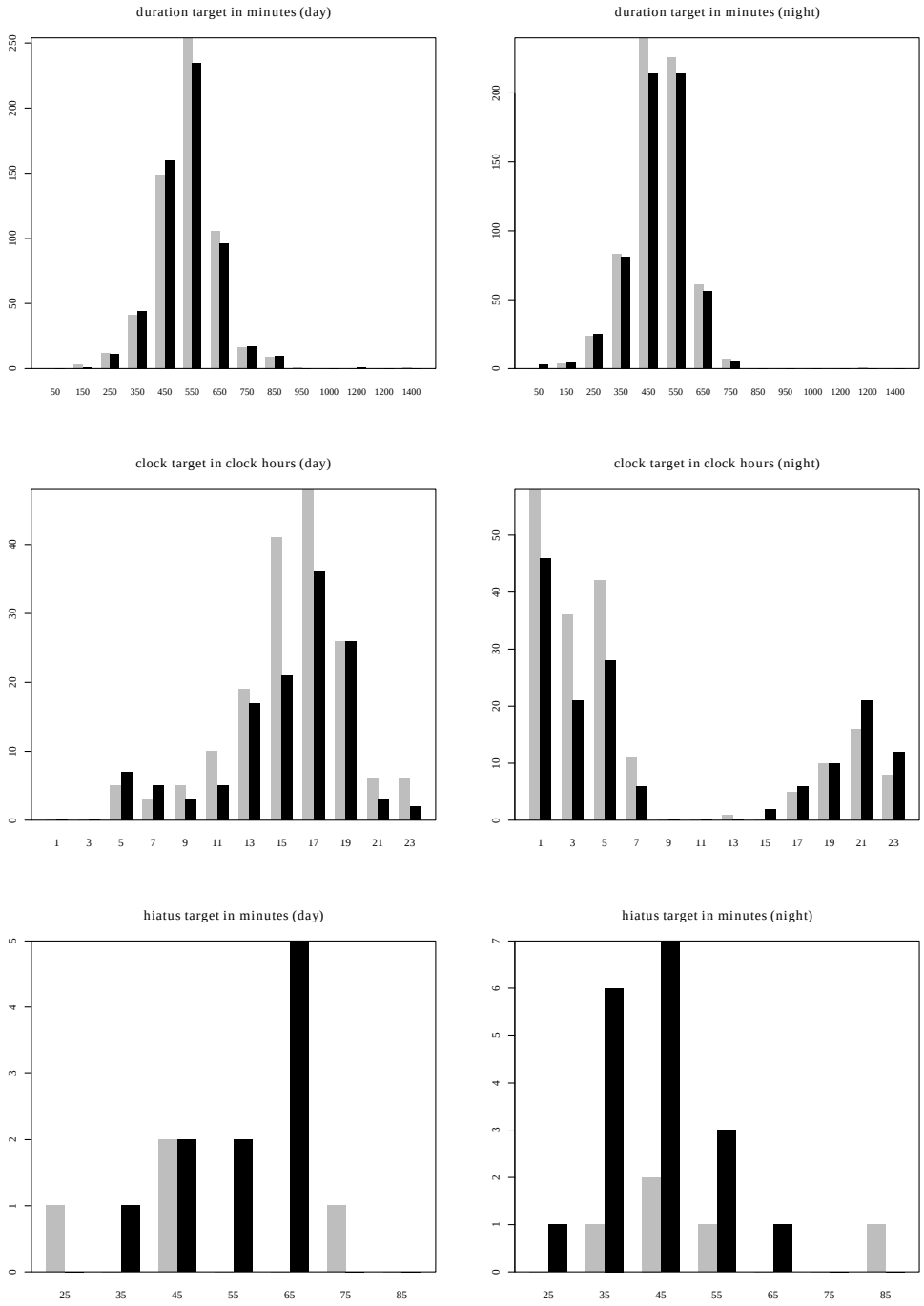

Figure B4: Histograms of median parameter estimates for old market (grey) and new market (black).

Appendix C Shift and Driver Classifications in Detail

Here, we offer more details on the classification of shifts and drivers. Table C1 gives an extended version of Table 4 and reports classifications stratified by shift type and market. Shift classification appear robust across all subsamples with only minor changes. Driver classifications were carried out separately for each of the four subsamples, implying that drivers may count multiple times if they fall into multiple samples. Unclassified drivers remain stable around 30 percent. Overall, the duration model accounts for slightly more drivers among night shifts than day shifts, but this conclusion does not hold for single drivers. Difference between the old and the new market are minor.

Table C1 also shows the absolute number of shifts consistent with their classification. These numbers are considerably lower than the overall absolute number of shifts best predicted by each of the six models, as shown in columns 2 and 5. This discrepancy is not surprising for two reasons. First, for each model, the total number of shifts include shifts of drivers that we could not classify. Second, there is no driver for whom all shifts are best predicted by the same model, so for each driver there are shifts inconsistent with the driver's classification. Indeed, Figure 2 illustrates that the first model rarely predicts more than 50 percent of shifts, indicating only a moderate level of consistency among drivers.

To assess consistency, we have added in parentheses to columns 4 and 7 the average of the share of consistent shifts among all shifts. Across all samples of drivers, consistency is around 40 percent for the goal models and between 30 and 35 percent for the utility models. There are several reasons for this finding. First, a larger number of models tested implies that — even by chance — percentages of individual models are reduced. Second, the nature of observational data implies that many personal constraints are unobserved, such as appointments or days with lack of motivation. Such constraints affect shift ends irrespective of the driver's strategy. Third, we hypothesize that many drivers use different strategies in parallel. These drivers predominantly use one strategy but change strategies in some regular fashion. Consider, for instance, a father who needs to pick up his children from sports every Tuesday and Friday at 6pm and is best predicted by the clock model for those days only.

If drivers use strategies in parallel, their second models deserve closer examination. Thus far, our most striking conclusion has been that the majority of drivers is best described by the earnings and duration models. Indeed, of all 1,193 day drivers in the old market we could classify, only 12 had neither of the two models among the first two, and of the 1,014 best predicted by either of the two models, 691 had the respective other as the second model. These proportions are similar for the other three samples, although only about half of the night drivers best predicted by either the earnings or duration model has both model among the first two[8]. Taken together, the first and second model account for about two thirds of the shifts completed by drivers on average.

Next we turn to overlaps between the four samples examined separately so far, beginning with the overlap between day and night drivers. In total, 1,153 drivers had sufficiently many day and night shifts to fall into both our samples and for 394 of them, day shifts are best predicted by the duration model. For 167 of these drivers, night shifts were also best predicted by the duration model with night parameter values on average

[8]Specifically, for day drivers in the new market/night drivers in the old market/night drivers in the new market, we find that 20/15/24 drivers out of 1,072/ 1,142/ 1,051 had neither the duration nor the earnings model among the first two and of the 929/935/866 drivers best predicted by either the earnings or duration model, 600/490/460 drivers had both models among their first two.

Table C1
Classifications of Shifts and Drivers

	All Drivers			Single Drivers only		
		Drivers			Drivers	
Model	Shifts	Count	Consistent	Shifts	Count	Consistent
Day Shifts in Old Market						
neoclassical utility	20,134 (8 %)	3 (0 %)	142 (34 %)	7,852 (9 %)	2 (0 %)	102 (35 %)
reference utility	18,410 (8 %)	3 (0 %)	66 (29 %)	6,766 (8 %)	1 (0 %)	16 (27 %)
earnings	68,523 (28 %)	422 (25 %)	22,625 (41 %)	26,370 (30 %)	148 (30 %)	9,488 (43 %)
duration	74,061 (31 %)	592 (35 %)	35,538 (42 %)	25,426 (29 %)	129 (26 %)	9,665 (39 %)
clock	46,793 (19 %)	169 (10 %)	8,563 (40 %)	16,536 (19 %)	40 (8 %)	2,789 (38 %)
hiatus	13,405 (6 %)	4 (0 %)	131 (38 %)	4,998 (6 %)	2 (0 %)	93 (39 %)
unclassified	—	505 (30 %)	—	—	176 (35 %)	—
total	241,326	1,698	67,065	87,948	498	22,153
Night Shifts in Old Market						
neoclassical utility	21,161 (11 %)	11 (1 %)	447 (32 %)	4,974 (11 %)	3 (1 %)	107 (32 %)
reference utility	14,866 (8 %)	4 (0 %)	115 (32 %)	3,228 (7 %)	2 (1 %)	65 (31 %)
earnings	46,372 (24 %)	289 (18 %)	12,582 (42 %)	12,442 (28 %)	101 (28 %)	5,001 (44 %)
duration	58,292 (31 %)	646 (40 %)	32,270 (40 %)	12,142 (27 %)	92 (25 %)	4,698 (37 %)
clock	38,674 (20 %)	187 (12 %)	7,414 (39 %)	9,283 (21 %)	47 (13 %)	2,136 (39 %)
hiatus	11,387 (6 %)	5 (0 %)	107 (34 %)	2,836 (6 %)	2 (1 %)	42 (34 %)
unclassified	—	473 (29 %)	—	—	114 (32 %)	—
total	190,752	1,615	52,935	44,905	361	12,049
Day Shifts in New Market						
neoclassical utility	17,872 (8 %)	7 (0 %)	441 (36 %)	7,826 (9 %)	2 (0 %)	145 (38 %)
reference utility	16,669 (8 %)	1 (0 %)	38 (31 %)	6,900 (8 %)	1 (0 %)	38 (31 %)
earnings	60,587 (27 %)	354 (23 %)	17,611 (41 %)	26,812 (29 %)	152 (28 %)	8,824 (42 %)
duration	70,124 (32 %)	575 (37 %)	34,791 (42 %)	26,744 (29 %)	143 (26 %)	9,994 (40 %)
clock	43,544 (20 %)	125 (8 %)	6,999 (41 %)	18,238 (20 %)	45 (8 %)	2,917 (40 %)
hiatus	12,671 (6 %)	10 (1 %)	455 (39 %)	5,355 (6 %)	3 (1 %)	253 (49 %)
unclassified	—	501 (32 %)	—	—	200 (37 %)	—
total	221,467	1,573	60,335	91,875	546	22,171
Night Shifts in New Market						
neoclassical utility	18,285 (11 %)	14 (1 %)	532 (31 %)	5,542 (11 %)	6 (2 %)	227 (31 %)
reference utility	13,866 (8 %)	1 (0 %)	51 (31 %)	3,509 (7 %)	0 (0 %)	0 (0 %)
earnings	41,148 (24 %)	262 (18 %)	10,725 (42 %)	13,535 (28 %)	101 (26 %)	5,183 (44 %)
duration	52,116 (31 %)	604 (41 %)	30,390 (40 %)	12,630 (26 %)	102 (26 %)	4,959 (38 %)
clock	33,625 (20 %)	152 (10 %)	5,989 (39 %)	9,919 (20 %)	59 (15 %)	2,583 (38 %)
hiatus	10,901 (6 %)	18 (1 %)	561 (36 %)	3,379 (7 %)	8 (2 %)	198 (34 %)
unclassified	—	413 (28 %)	—	—	118 (30 %)	—
total	169,941	1,464	48,248	48,514	394	13,150

6 percent below their day counterparts. In contrast, for 58 drivers, night shifts were best predicted by the earnings model and for 112 drivers, night shifts could not be classified.

Consider next the overlap between drivers in the old and new markets. In total, 1,418 drivers had sufficiently many shifts in the old and new market to fall into both of our samples. For 632 of them, shifts in the old market were best predicted by the duration model and for 401 of those, shifts across both markets were best predicted by the duration model. On average, parameter values for the new market were 0.2 percent lower than their counterparts in the old market. In contrast, for 52 drivers, night shifts were best predicted by the earnings model and for 154 drivers night shift could not be classified.

Appendix D Aggregate Outcomes in Detail

Table D1 is an extension of Table 6 and reports elasticity estimates and mean earnings stratified by shift type and market. Differences between subsamples are large for the utility and hiatus models that account for very few drivers only. For the three most predictive models, median elasticity estimates of drivers are fairly stable: The median for drivers best predicted by the earnings target consistently falls between −0.70 and −0.86, whereas it fluctuates around zero for the duration and clock models. These conclusions also hold for elasticity estimates across shifts. For the earnings model, we find a moderate increase from the old market to the new and for the duration model, we find slightly more positive elasticities for night than for day shifts but they remain close to zero. Estimates for the utility and the hiatus models fluctuate considerably, strengthening our suspicion that these results are strongly affected by between-driver variance.

Table D1
Elasticities And Hourly Earnings by Model

	Across Drivers				Across Shifts			
			Earnings				Earnings	
Model	Drivers	Elasticity	Mean	SD	Shifts	Elasticity	Mean	SD
Day Shifts in Old Market								
neoclassical utility	3	1.07	15.8	1.6	20,134	−0.31	15.8	6.9
reference utility	3	−0.01	14.6	5.5	18,410	−0.37	16.4	6.3
earnings target	422	−0.70	15.1	5.2	68,523	−0.70	15.8	6.2
duration target	592	−0.00	15.9	4.7	74,061	−0.03	15.7	6.6
clock target	169	−0.04	18.2	5.0	46,793	0.03	16.1	6.4
hiatus target	4	0.84	17.6	4.3	13,405	0.05	14.8	7.5
Night Shifts in Old Market								
neoclassical utility	11	1.29	24.0	3.1	21,161	0.25	19.5	7.6
reference utility	4	−1.77	14.0	3.2	14,866	−0.06	19.8	7.3
earnings target	289	−0.72	13.8	5.5	46,372	−0.67	17.2	7.2
duration target	646	0.10	20.5	4.8	58,292	0.12	19.5	8.3
clock target	187	0.28	20.0	5.0	38,674	0.35	18.9	7.4
hiatus target	5	1.19	18.0	6.2	11,387	0.61	16.4	7.8
Day Shifts in New Market								
neoclassical utility	7	2.09	22.4	5.8	17,872	0.03	18.2	7.4
reference utility	1	−0.52	24.7	—	16,669	−0.35	18.6	6.8
earnings target	354	−0.83	16.6	6.0	60,587	−0.97	17.4	6.7
duration target	575	−0.01	18.5	4.8	70,124	0.01	17.9	6.9
clock target	125	0.07	20.6	5.8	43,544	0.14	17.9	6.8
hiatus target	10	−1.12	17.3	4.0	12,671	0.01	16.3	7.5
Night Shifts in New Market								
neoclassical utility	14	0.92	23.7	6.6	18,285	0.35	21.7	8.3
reference utility	1	−0.40	21.8	—	13,866	−0.02	22.6	8.0
earnings target	262	−0.86	17.2	7.1	41,148	−0.99	19.6	8.0
duration target	604	0.10	23.1	5.3	52,116	0.14	22.3	8.9
clock target	152	0.02	21.0	6.0	33,625	0.33	21.1	8.2
hiatus target	18	−0.02	20.6	5.3	10,901	0.19	18.7	13.9

Notes: Across drivers: hourly earnings averaged across consistent shifts and drivers, hourly earnings SD gives SD across drivers, elasticity gives median elasticity estimate from IV regression for each driver; across shifts: hourly earnings averaged across shifts of the same strategy, elasticity obtained from regression using all shifts of the same strategy; for brevity, no separate display for single drivers.

Table D1 also reports mean hourly earnings across the four subsamples. Overall, hourly earnings tend to be higher during night shifts than day shifts. Nonetheless, we find only minor differences between mean earnings of different models — between day and night shifts and between the old and the new market. As before, the earnings models consistently exhibits the lowest mean hourly earnings among the three best predicting models, sometimes the lowest overall. Calculated across shifts, the utility models exhibit somewhat higher earnings than other models but standard deviations remain large. For a better illustration of differences, Figure D1 shows the kernel density estimates of the distribution of earnings across drivers predicted by the earnings and duration models in grey, as well as the utility model with highest mean earnings in black. Across all four subsamples, differences between the models are existent but appear small in comparison to the variance within each group of drivers.

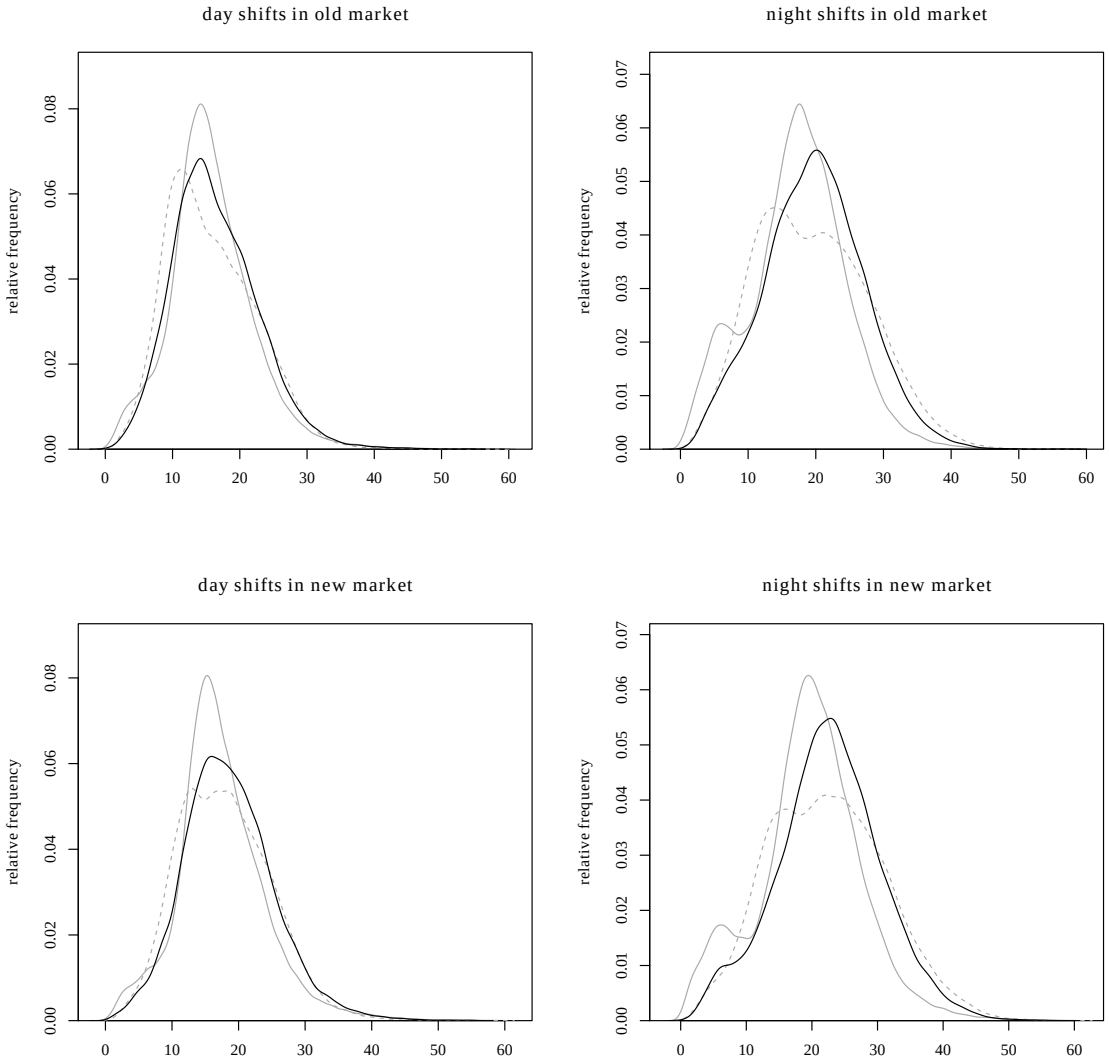

Figure D1: Plots of kernel density estimates for distribution of mean hourly shift earnings for earnings model (grey, solid), duration model (grey, dashed), and the utility model with highest mean earnings (black); includes all shifts, rare outliers above € 60 excluded.