Designing Promises with Reference-Dependent Customers: The Case of Online Grocery Delivery Time

Click for Latest Version

(Preliminary Draft)

George Gui[*] Tilman Drerup[†]

Dec 9, 2022

Abstract

Firms often need to promise a certain level of service quality to attract customers, and a central question is how to design promises to balance the trade-off between customer acquisition and customer retention. For example, most E-commerce platforms need to promise a certain delivery time. Over-promising may attract more customers now, but its impact on future retention depends on consumer inertia, learning, and loss aversion. Empirical analysis of this topic is challenging because the realized and promised service qualities are often unobserved or lack exogenous variation. We leverage a novel dataset from Instacart that directly observes variation in promised and actual delivery time to study this problem. We apply a generalized propensity score method to nonparametrically estimate the impact of delivery time on customer retention. Consistent with reference dependence and loss aversion, we document that customers are around 92% more responsive once the delivery becomes late. Our results inform a structural model of learning and reference dependence that illustrates the importance of estimating loss aversion and distinguishing promise-based reference points from expectation-based reference points: the company would forgo millions of dollars in revenue if it underestimates loss aversion or assumes expectation-based reference points.

[*] Graduate School of Business, Stanford University, ggui@stanford.edu

[†] Instacart, tilman.drerup@instacart.com

1 Introduction

Firms often need to promise a certain level of service or product quality to attract customers, and a key question is how firms should design this promise. For example, E-commerce platforms, such as Amazon, Uber, and Instacart, all need to promise a delivery time. One possible strategy is "under-promise and over-deliver" (Peters (1988), Sewell and Brown (2009), Forbes (2013)). Customers may use promises as reference points (Kahneman and Tversky (1979)), becoming more satisfied when the actual quality exceeds the under-promised quality. Another strategy is to over-promise now to acquire more customers (Spence (1977)). These customers may be more willing to purchase the product or service in the future due to inertia or increased familiarity (Klemperer (1995)). A key managerial input is how customers evaluate their past promises for making future decisions. Without empirically assessing how customers respond to promises, firms cannot design better promise strategies to balance the trade-off between customer acquisition and customer retention.

Empirical analysis of this topic has been challenging due to data limitation: objective and quantifiable measures[1] of promised and actual service qualities, coupled with actual purchase data, are rarely observed. This data limitation prevents researchers from empirically assessing impacts of promised and actual qualities on customer purchase. Specifically, what are the economic consequences when the actual quality deviates from the promised quality? Are customers more responsive when promises are broken? Is it managerially important to account for this potentially asymmetric response? How do customers respond differently as they gain more experience? How should firms design promises accordingly?

Our paper develops an empirical framework to design promises in the context of online grocery delivery. We leverage a novel dataset from Instacart, an online platform that offers grocery delivery services to North American customers. We observe a panel data of customer order history that includes more than 10 million orders in a 9-month period. For each order, we observe both promised and actual delivery time, as well as how much customers spend before and after experiencing these delivery times. Because we observe direct measures of promised and actual qualities in terms of delivery time, we can assess the impact of promises on customer retention and study its implication for designing promises.

We use a generalized propensity score method (Imai and van Dyk (2004) and Hirano and Imbens (2005)) that leverages detailed order characteristics to nonparametrically estimate the impact of delivery time on customer retention. Our estimation accounts for several empirical

[1] Service qualities are often measured using survey data based on customer recall (e.g., Parasuraman et al. (1985) and Parasuraman et al. (1988)).

challenges, including that the delivery time is an endogenous and continuous treatment, and preference for delivery time may be nonlinear and heterogeneous across customers.

We document several empirical patterns on how delivery time affects retention. First, both over-delivering and under-delivering affect customer retention, suggesting customers are attentive to the actual delivery time.[2] Second, customers become 92% more responsive to delivery time once it passes promises, suggesting that customers use promises as reference points. Third, this pattern holds for experienced customers who used to receive delivery much earlier than promises, suggesting that past experience has limited impact on reference point formation. Fourth, past experience has a significant impact on the degree of responsiveness: the impact of the most recent delivery experience on customer retention diminishes as customers gain more experience, suggesting evidence of learning.

Based on these empirical findings, we develop a model that informs the design of promises. The model accounts for customers using promises as reference points and being more responsive to losses than to gains. The model also accounts for learning, allowing customers to update their beliefs about promises as they gain more experience. The model is capable of simulating counterfactuals under different promises, enabling firms to choose the optimal strategy.

To demonstrate the credibility of the simulated counterfactuals, we conduct an out-of-policy test to validate model predictions, leveraging both an A/B test[3] and a policy change. We show that even if the model is estimated based on a training dataset that has no variation in promises, it is capable of successfully predicting what will happen when promises change during the A/B test and the policy change, demonstrating the external validity of our model.

We use the validated model to illustrate the importance of identifying reference points and loss aversion for designing promises. If the platform underestimates loss aversion, or assumes that customers use the average delivery time rather than the promise as the reference point, the platform will set promises that are too aggressive, leading to a high probability of late deliveries. These suboptimal promise strategies translate to 10.69 to 76.27 million dollars loss in annual revenue. Because improving promises is almost costless,[4] and online grocery is a billion-dollar industry with a high customer acquisition cost, this represents an economically

[2] This empirical finding is different from lab studies conducted by Gneezy and Epley (2014) that document the benefit of exceeding promises is limited.

[3] The A/B test comes from a later period that was designed to measure the short-term impact of promises on customer acquisition. The A/B test alone is insufficient for designing promises because it does not control for the actual delivery time: randomizing promises will indirectly change actual delivery time through a queuing algorithm.

[4] We focus on optimizing promises while holding the actual delivery time the same, such that the fulfillment cost remains similar.

significant difference in profit.

The remainder of the paper is organized as follows: Section 2 reviews the literature. Section 3 illustrates a theoretical framework for designing promises. Section 4 describes the data. Section 5 causally estimates the impact of delivery time on customer retention and demonstrate that the result is consistent with reference dependence and loss aversion. Section 6 rules out several alternative explanations. Section 7 discusses how the estimated impact is related to customer experience and documents evidence of learning. Section 8 builds a structural model that rationalizes the empirical findings. Section 9 validates the model using a policy change and an A/B test. Section 10 simulates counterfactual policies to demonstrate the value of identifying reference points and loss aversion in designing promises. Section 11 concludes with questions for further study.

2 Literature

Our paper contributes to the literature on customer retention by empirically studying promises. The existing literature has focused on how retention is causally affected by the actual quality experienced by customers (Bolton (1998), Braun and Schweidel (2011), Sriram et al. (2015), Ascarza et al. (2018), Kim (2021)), which help answer questions about how to improve the actual products or services. Our paper asks a different question: given the actual products or services firms are selling, how should firms design promises? The answer to this question partly depends on how the actual quality interacts with promised quality to affect customer retention. To the best of our knowledge, we are the first to causally estimate the economic consequences of over-delivering versus under-delivering by combining objective measures of service qualities with actual retention data.

Our empirical findings are related to the literature on reference point formation that is theoretically ambivalent. Kahneman and Tversky (1979) take the status-quo as reference points but also suggest expectations may be reference points.[5] Subsequent theories propose reference points could be determined by rational expectations (Kőszegi and Rabin (2006)). The theory of expectation-based reference points has been examined in several lab studies (Marzilli Ericson and Fuster (2011), Gill and Prowse (2012), Heffetz and List (2014)) and has been documented using car prices (Huang and Liu (2021)) and eBay auction outcomes (Backus et al. (2021)). Our paper studies online grocery delivery and documents that past experience has limited impact

[5] According to Kahneman and Tversky (1979), "So far ... the reference point was taken to be the status quo, or one's current assets. Although this is probably true for most choice problems, there are situations in which gains and losses are coded relative to an expectation or aspiration level that differs from the status quo"

on reference point formation compared to promises.

Our empirical findings are related to the marketing models (e.g., Bolton (1998), Ho and Zheng (2004), Kopalle and Lehmann (2006)) that incorporate prospect theory or disconfirmation, which require assuming a certain reference point. We highlight the importance of empirically assessing assumptions about reference point formation. We distinguish expectation – an internal belief that depends on customer prior experience – from promise – a salient and explicit message that firms communicate to customers. We illustrate that assuming rational expectations to be reference points may lead to sub-optimal promise policies.

Our model is related to the theoretical literature on expectation management. Ho and Zheng (2004) consider a static demand model with competition and does not consider dynamics and repeated purchases. Joshi and Musalem (2012) and Martin and Shelegia (2021) consider models where dissatisfied customers penalize firms through word-of-mouth and does not consider repurchase decisions. Kopalle and Lehmann (2006) consider a two-period theoretical model where firms have no uncertainty over their own product quality. We consider a multi-period model that is useful for designing promises when customers make repeated purchases and firms have uncertainty over their own service qualities, which can be applied to most E-commerce platforms.

3 Theoretical Framework

One plausible framework is to assume that customers are reference-dependent and loss averse: customers evaluate their experience based on a certain reference point, and become disappointed if the realized quality is worse than this reference point. Although researchers generally agree that losses loom larger than gains, a fundamental question is what is the degree of loss aversion and what is perceived as a loss (Barberis (2013)). We present a simple theoretical framework to demonstrate why empirically identifying reference points and loss aversion is essential for designing promises. For illustrative purpose, we first consider a simple two-period setting with homogeneous customers, and later extend the model to a multi-period setting in the empirical section.

Consider an E-commerce platform that needs to promise customers with a delivery time before customers make their purchase decisions. The platform knows that promising these customers with faster delivery time $Promise_1$ can increase purchase Y_1 in period 1.[6] The platform is uncertain about the precise delivery time D required to fulfill the order, but knows

[6] This knowledge can be easily obtained by running A/B test that gives customers different promised delivery time and compare the conversion rate

that this delivery time follows a known probability distribution F. The question is how the platform should set promises given the knowledge of the distribution F.

The answer depends on how customers respond in the second period. The related literature on customer satisfaction and retention (e.g. Kumar et al. (1997), Bolton (1998), Kopalle and Lehmann (2001), Kopalle and Lehmann (2006), Gijsenberg et al. (2015)) typically starts with the assumption that losses loom large than gains. Customers evaluate the delivery time they experienced, D_1, relative to a certain reference point, r_1, and become disappointed if the delivery arrives later than the reference point. This reference point hence will affect their subsequent purchase decision Y_2:

$$Y_2 = \begin{cases} \alpha - \beta(D_1 - r_1) & D_1 \leq r_1 \\ \alpha - (\beta + \gamma)(D_1 - r_1) & D_1 > r_1 \end{cases} \tag{1}$$

where $\gamma \geq 0$ characterizes the degree of loss aversion or disappointment aversion when the delivery arrives later than the reference point. These disappointed customers may perceive the delivery time to be longer than they actually are or become less loyal to the platform, thus less likely to repurchase.

A promise that maximizes the total purchase should balance customer acquisition and customer retention, such that:

$$\underbrace{-\frac{\partial Y_1}{\partial Promise_1}}_{\substack{\text{Marginal impact} \\ \text{on customer acquisition}}} = \underbrace{\frac{\partial \bar{Y}_2}{\partial Promise_1}}_{\substack{\text{Marginal impact} \\ \text{on customer retention}}} \tag{2}$$

where $\bar{Y}_2$ is the expected purchase in the second period given the distribution of delivery time:

$$\bar{Y}_2 = \alpha - \beta E[D_1 - r_1] - \gamma E[(D_1 - r_1)I(D_1 > r_1)], \tag{3}$$

which implies that the impact of reference point on retention is determined by the degree of loss aversion γ and the probability of disappointment $1 - F(r_1)$:

$$\frac{\partial \bar{Y}_2}{\partial r_1} = \beta + \gamma \left[1 - F(r_1)\right] \tag{4}$$

We demonstrate that setting the optimal promise requires identifying the value of the reference point given a certain promise. If customers usually receive their delivery θ minutes earlier than promises, one possibility is that these customers use the expected delivery time, $Promise_1 - \theta$, as the reference point. Another possibility is that these customers use the conservative promise itself as the reference point, making them less likely to be disappointed.

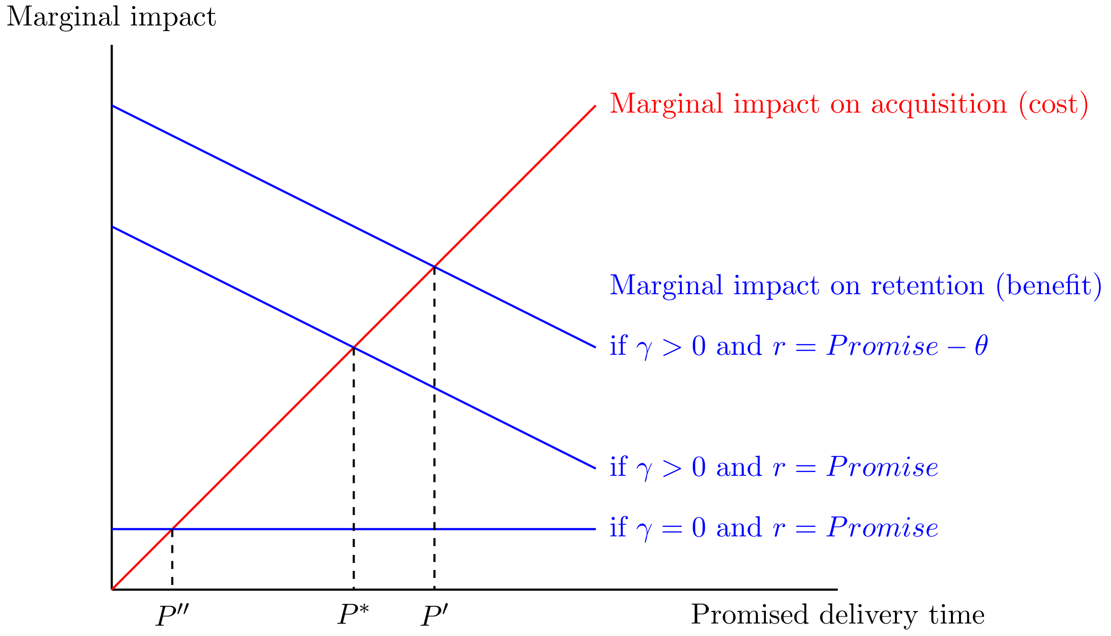
Equation 4 implies that this lower probability of disappointment decreases the marginal impact of promises on retention:

$$\begin{aligned}
&\frac{\partial \bar{Y}_2}{\partial Promise_1}\Big|_{r=Promise_1} - \frac{\partial \bar{Y}_2}{\partial Promise_1}\Big|_{r=Promise_1-\theta} \\
=&\gamma \left\{[1 - F(Promise_1)] - [1 - F(Promise_1 - \theta)]\right\} \\
=&\gamma \left[F(Promise_1 - \theta) - F(Promise_1)\right] < 0
\end{aligned} \tag{5}$$

Because of this distinction, correctly specifying reference point and loss aversion is crucial for setting promises. Figure 1 illustrates how the optimal promises differ given different assumptions about loss aversion and reference formation. If firms incorrectly assume that customers are not loss-averse, they may set promises that are too ambitious ($P'' < P^*$) because they underestimate the negative impact of disappointing customers.

Even if firms know that customers are loss-averse and promises affect reference points, firms may set sub-optimal promises: if firms incorrectly assume that customers use $r = Promise_1 - \theta$ as reference points when customers actually use promises as reference points $r = Promise_1$, firms may set promises that are too conservative than optimal ($P' > P^*$). Therefore, although it is intuitive to assume that promises *affect* reference points, the assumption is not precise enough.[7] For designing optimal promises, firms not only need to estimate the degree of loss aversion but also understand the exact mapping between promises and reference points.

Figure 1: Optimal promises under different reference points and loss aversion


Marginal impact
Marginal impact on acquisition (cost)
Marginal impact on retention (benefit)
if $\gamma > 0$ and $r = Promise - \theta$
if $\gamma > 0$ and $r = Promise$
if $\gamma = 0$ and $r = Promise$
P''
P^*
P'
Promised delivery time

To design promises given *any* distribution F of the realized quality, this theoretical frame-

[7] For $r = Promise$ and $r = Promise - \theta$, promises have the same marginal effect on reference point, but they generate different optimal promises.

work suggest that it is crucial to estimate 1) how promises affect short-term customer acquisition and 2) how promises interact with the realized quality to affect long-term customer retention. Although it is straightforward for online platforms to randomize promises to measure their impact on customer acquisition, it is challenging to measure how promises interact with the realized quality to affect customer retention because platforms typically have imperfect control over the realized quality.[8] The rest of the paper focuses on tackling this empirical challenge on customer retention, and then briefly discuss an A/B test on customer acquisition in Section 9.2.

4 Data

We leverage a dataset from Instacart, an online grocery platform that offers grocery delivery services to North American customers. The platform partners with grocery stores and enables customers to shop from these grocery stores through its website or mobile app. When customers check out, they are given several delivery options. Throughout most of our study, the platform calculated the promised delivery time based on the order characteristics and the expected supply condition, and then round the earliest delivery window to 30 minute, 1-hour, 2-hour, or 5-hour. Figure 2 gives an example of the immediate delivery options customer see on the checkout page. Because the majority of deliveries in our dataset are promised to arrive **within 2 hours**, we focus on this subset of data in our main analysis.

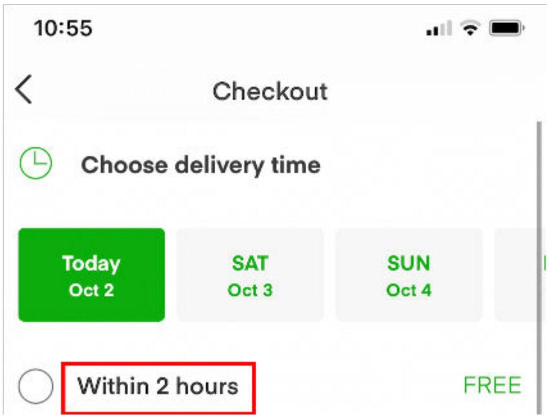


Figure 2: Delivery options displayed to customers when they are about to check out

We observe over 10 million 2-hour delivery orders placed in a 9-month window. For each order made by customers, we observe both promised and realized delivery times, which serve as measures of service quality. We also observe customer ratings and future spending after the order is fulfilled, which we use to construct measures of customer satisfaction and retention.

[8] For example, the delivery time can be affected by unexpected traffic.

The dataset also includes other order characteristics known by customers when they place the order, such as basket size, time of the order, and store locations.

4.1 Variation in delivery time

A key feature in our data is the variation of the realized delivery time at minute level. Figure 3 illustrates an example of the distribution of delivery time in a given month. Thanks to this variation in delivery time, we observe many orders that have exactly the same delivery minute for any delivery minute D. We focus on customer response when the delivery minutes are between 60 to 150. For any minute D in this range, we observe a minimum of $10,000$ orders that take exactly D minutes to arrive.

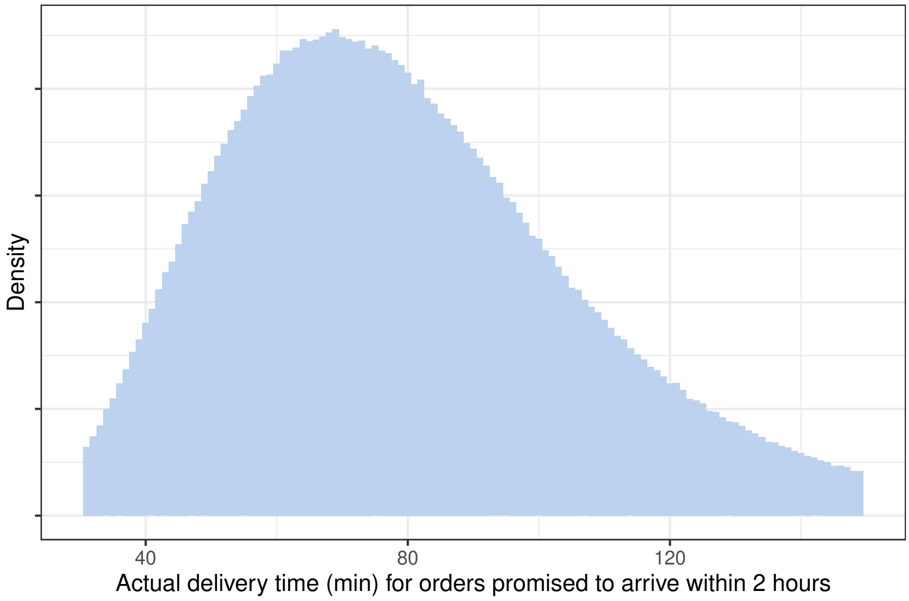


Figure 3: An example of the distribution of delivery time in a month

Our data is novel because it has direct measures of the promised and the realized quality. Because the realized quality is a continuous measure that rarely coincides with the promised quality, we can quantify both positive and negative deviation from promises. Importantly, we observe customer actual purchase decisions, allowing us to measure customer response based on revealed preference and studying the economic consequences of over-delivering versus under-delivering.

4.2 Measures of retention and satisfaction

We construct measures of customer retention based on how much customer spend in the future after a given delivery. To account for households having different baseline expenditure level, we measure the percentage change in customer expenditure by taking the log difference between how much customers spend in the next 4 weeks after a delivery versus how much customers spend in the previous 4 weeks: $log(\text{FutureExpenditure}) - log(\text{PastExpenditure})$.[9]

To supplement our main analysis, we also construct measures of customer satisfaction based on self-reported rating. After each delivery order is completed, customers are asked to rate their experience on a scale of 1-5 stars. One common challenge with rating data is that not all customers report their rating. To account for this missing data issue, we focus on whether customers explicitly give a five star rating. Intuitively, if customers are more satisfied, they are more likely to explicitly give a five-star rating. If customers are less satisfied, they are more likely to give sub-five-star ratings or not give ratings. We consider whether explicitly giving a five-star as a binary measure of customer satisfaction that is not subject to the missing data issue.

5 Empirical Framework

As noted in Section 3, testing whether customers use promises as reference point and estimating the degree of loss aversion is crucial for designing promises. This section presents nonparametric evidence that is consistent with promised-based reference points and loss aversion.

5.1 Objectives

Our empirical strategy is to estimate a continuous causal response curve that characterizes the impacts of delivery time on retention: how would customer retention be affected if the delivery time were D. To test whether reference points are around promises, we leverage the property that customer response is kinked around the reference point: customers become more responsive to delivery time once it passes the reference point.

Formally, we follow a generalized version of the potential outcome framework that accounts for continuous treatment (Hirano and Imbens (2005)). Using this continuous framework is necessary because we are interested in the shape of the response curve. If we convert the continuous treatment – delivery time – into a binary treatment – whether the delivery is late, we can no longer identify the reference point nor loss aversion. Let $Y_i(D)$ be the potential retention if order i were to have a delivery time of D minutes. We define the causal response curve μ as:

[9] We add 1 to the expenditure to account for cases when expenditure is 0.

Definition 1. *(Causal Response Curve)*

$$\mu(D) \equiv E[Y_i(D)],$$

which summarizes the average potential retention when delivery time is D. This causal response curve can be used to answer counterfactual questions such as what would happen if all orders in the sample were to have a delivery time of D. The curve also enables us to evaluate the impact of over-delivering and under-delivering. For example, what is the benefit of arriving 30-minutes early vs 10-minutes early? What is the cost of arriving 10-minutes late vs 30-minutes late?

5.2 Empirical challenges

5.2.1 Endogeneity

One empirical challenge to estimate the causal response curve is endogeneity. Delivery times are not random. They are affected by factors including customer locations, order time, and the number and types of items that customers order. Table 6 illustrates that on-time and late orders differ in several dimensions. Customers who receive their orders on time tend to live closer to the grocery store, place fewer items per order, and order less often on weekends. Because of this endogenous delivery time, we should not directly compare retention between customers who experience late deliveries versus customers who experience on-time deliveries.

To address this endogeneity issue, we leverage covariates that capture order-level characteristics as well as customer order history at the time when the order is placed. These covariates, denoted as X_i, include number of items in the basket, time of the order and distances from stores, customer locations and distances from stores, past average delivery time, customer experience, and past order frequency. Our identification strategy relies on the assumption that delivery time and potential customer retention are independent after conditioning on these observed covariates:

Assumption 1. *(Conditional Independence)*

$$Y_i(D) \perp D_i | X_i$$

where $Y_i(D)$ is the potential customer retention if the delivery were to take D minutes, D_i is the actual delivery time, and X_i are covariates that capture the order and customer characteristics.

Intuitively, customers who place similar orders have similar preferences. This similarity in preference will hold in the future unless these similar customers experience different service qualities. Therefore, we can compare these similar customers that experience different delivery time to estimate the causal response curve.

5.2.2 Non-linearity and heterogeneity

Another challenge is that the impact of delivery time on retention may be non-linear. One tempting method is to run a linear regression between retention Y_i and delivery time D_i, while controlling for other covariates X_i. Although this method is easy to implement, and can capture the first-order impact of delivery time, that longer delivery time decreases retention, the method is too restrictive to uncover more granular patterns in the data. If customers have pre-determined plans, then the impact of delivery time may be discontinuous and step-wise because customers will be extremely disappointed once the delivery arrives later than plans. If customers are loss averse, then the preference will be kinked around reference points. Imposing parametric assumptions on how the delivery time affects retention at the estimation stage will prevent researchers from evaluating whether such assumptions are plausible compared to alternative specifications.

A related issue with linear regression is that customers have heterogeneous preferences for time. Flexibly accounting for consumer heterogeneity is important because the measurement of loss aversion is confounded with consumer heterogeneity (Bell and Lattin (2000)): even if customers are not loss-averse, researchers may find the relationship between delivery time and retention to be asymmetric because of heterogeneity. Similarly, even if customers are loss-averse, researchers may find the relationship between delivery time and retention to be linear. Because linear regression with additive controls assume customers have homogeneous preference for the treatment, they are not suitable for exploring the shape of the response curves.[10]

5.3 Method: Generalized propensity score

We use the generalized propensity score method (Imai and van Dyk (2004) and Hirano and Imbens (2005)) to operationalize this estimation. The method allows us to incorporate Assumption 1 to flexibly control for customer heterogeneity. The method also does not impose any parametric assumption on the shape of the response curve in the estimation stage, allowing us to later assess whether a class of parametric assumptions motivated by theories are plausible.

Following the method, we first estimate the probability density function of delivery time $\widehat{f}(D|X)$ that depends on order characteristics X. Then for each order i with delivery time D_i, we derive a generalized propensity score $\widehat{f}(D_i|X_i)$ based on the estimated density. This generalized propensity score captures the propensity that the order with characteristics X_i will have a delivery time of D_i, which enables us to use the inverse propensity score weighting

[10] Similarly, binscatter regressions (Cattaneo et al. (2021)) also assumes additive separability and is therefore not suitable for analyzing this problem if customers are heterogeneous.

approach to estimate the average potential outcome at any given delivery minute D:

$$\widehat{\mu}(D) = \frac{\sum_{i:D_i=D} \widehat{w}_i Y_i}{\sum_{i:D_i=D} \widehat{w}_i}$$

where the weight w_i is the inverse of the generalized propensity score $\widehat{w}_i = \frac{1}{\widehat{f}(D_i|X_i)}$. Intuitively, the weight is needed because some orders have higher propensity to arrive around minute D than others. This difference in propensity implies that orders with a particular delivery time D do not constitute a representative sample. For example, orders that arrive early tend to have below-average basket size. By including the weight, we account for this difference in propensity and make the weighted sample representative. Appendix A discusses the detailed implementation of this method.

Because our analysis focuses on deliveries promised to arrive "**within 2 hours**", we estimate the impact of deviating from promises by summarizing

$$\widehat{\mu}(D) - \widehat{\mu}(120)$$

5.4 Impacts of deviating from promises

Figure 4 summarizes our estimation results conducted for all orders that have a promised delivery time of 120 minutes. Each point represents the causal impact of deviating from promises at minute D: $\widehat{\mu}(D) - \widehat{\mu}(120)$. We obtain 91 such point estimates, each point representing a minute between 60 minute to 150 minute.[11] These estimators suggest that customers do respond to past delivery time, responding positively to early arrivals and negatively to late arrivals. To visualize the difference between customers responsiveness to positive versus negative deviation, we use estimated points to run linear regression after partitioning the points based on whether it is on-time or late. The dashed line, fitted by extrapolating the on-time delivery data, suggest that customers become more responsive to delivery time when it is late.

5.5 Reference point estimation

In Figure 4, the fitted *lines* only demonstrate one possible parametric specification to fit the *points* that are nonparametrically estimated. Because these *points* are estimated without any shape restrictions, we can assess whether alternative parametric specifications motivated by prospect theory are plausible, without imposing the assumption that reference points must be around promises.

Assume that the potential retention follows a piece-wise linear function with an unknown reference point:

[11] We focus on these range because the number of observations outside of this range is limited.

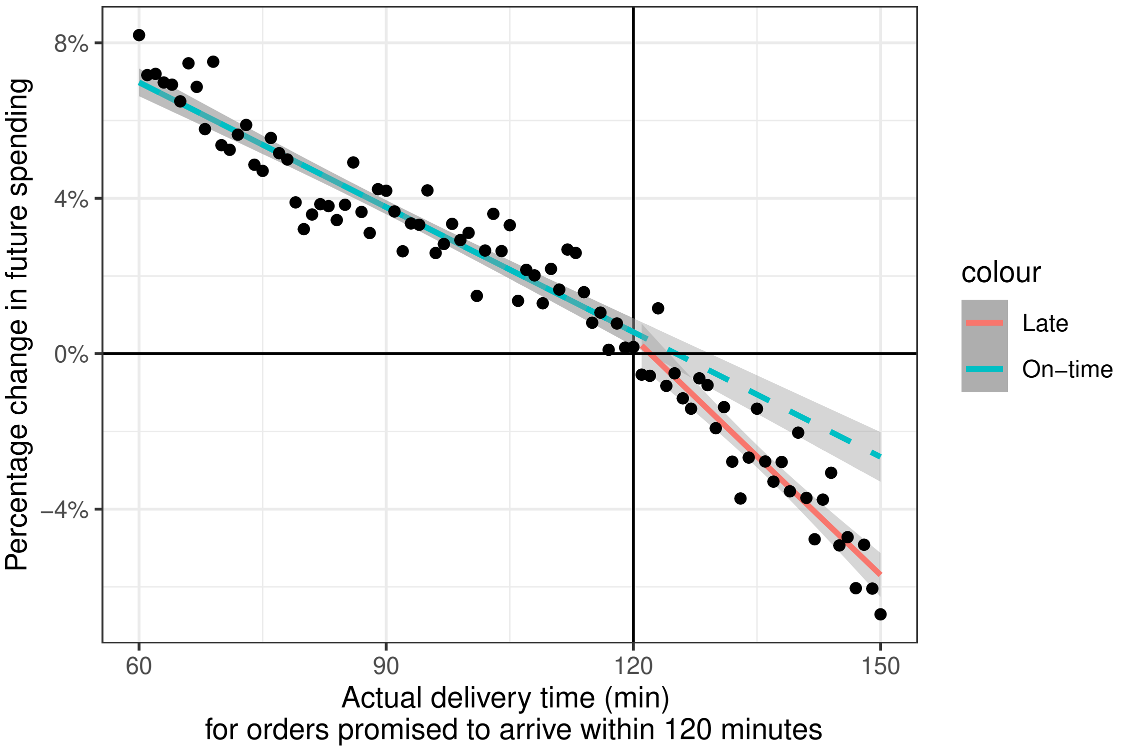


Figure 4: Impact of delivery time on customer retention, estimated using generalized propensity score method

$$Y_i(D) = \begin{cases} \alpha_i + \beta_i(D - r) & D \le r \\ \alpha_i + (\beta_i + \gamma_i)(D - r) & D > r \end{cases} \tag{6}$$

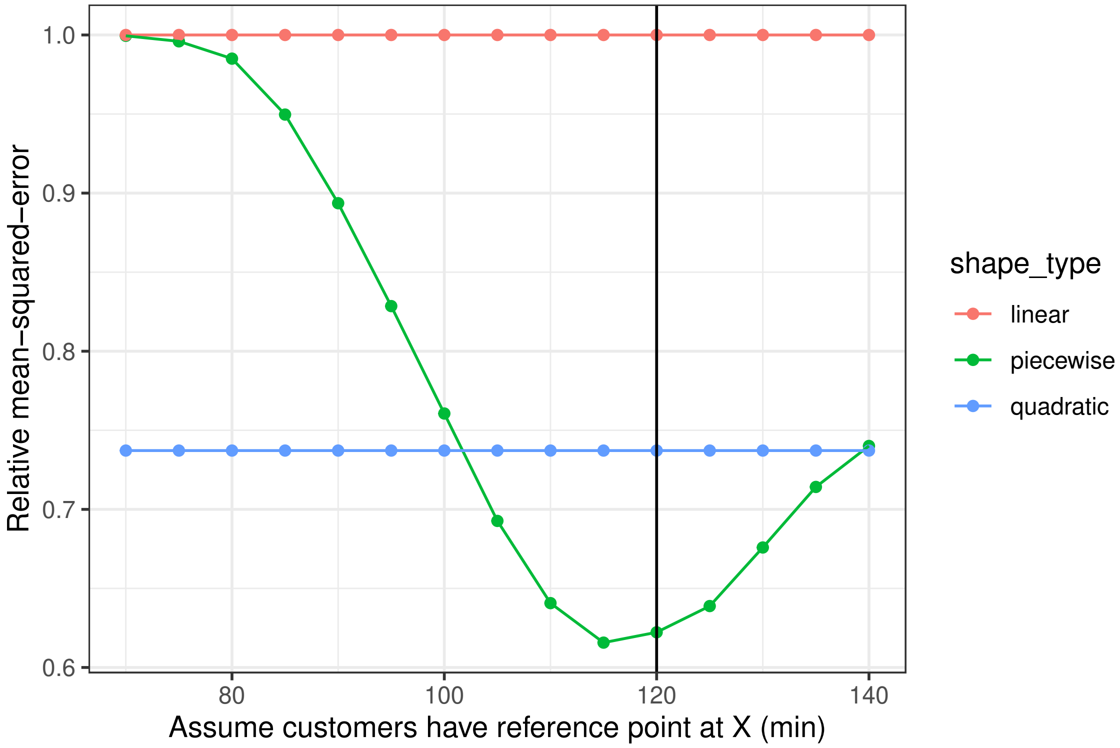
where β_i characterizes the responsiveness to delivery time when the delivery arrives before the reference point, and γ_i captures loss aversion. The average response curve across all customers should then also follow a piece-wise linear pattern:

$$\mu(D) = E[Y_i(D)] = \begin{cases} \alpha + \beta(D - r) & D \le r \\ \alpha + (\beta + \gamma)(D - r) & D > r \end{cases}$$

where $\beta = E[\beta_i]$ summarizes the average responsiveness to delivery time and $\gamma = E[\gamma_i]$ summarizes the loss aversion. When the reference point r is known, we can directly estimate (β, γ) using standard linear regression. However, because prospect theory is ambivalent over how reference point is formed (Barberis (2013)), and Section 3 has illustrated that the misspecification of reference point can lead to suboptimal promises, we treat the reference point r as unknown and directly estimate it.

We estimate the reference point by running piece-wise linear regressions with different assumed reference points. Intuitively, if the reference point is correctly specified, then a piece-wise linear regression with the correctly assumed reference point should outperform similar regressions with incorrectly assumed reference points. Figure 5 compares the goodness-of-fit under different assumed reference points. The model is best fitted when the reference point is assumed to be around promises, outperforming the linear model by more than 30% in mean-squared-error.

Figure 5: Customer response to delivery time are best approximated by a piece-wise linear function that changes slope around promises.



Notes: This relative mean-squared error is calculated based on the mean-squared-error when we assume a certain reference point versus when we assume a linear model that does not have a reference point.

Figure 5 also evaluates whether the asymmetric response can be better approximated by a quadratic function. This quadratic shape could be driven by non-linear preference in delivery time that is independent of reference point. The piece-wise linear model still outperforms the quadratic model, suggesting that the pattern is better explained by customers using promises as reference points.

Table 1 compares the regression results when using promises as reference points versus when assuming customer response is linear or quadratic. The piece-wise linear model has higher R^2 than the other specifications. Its high R^2 of 0.96 also suggests that this piecewise parametric specification is a good approximation for customer response. The model estimates indicate that customers are around 92% more responsive to delivery time when it is late: arriving an hour early increases the future spending by around 6.5%, but arriving an hour late decreases the future spending by 12.5%.

Table 1: Effect of delivery time on retention under different shape restriction

	Outcome: Percentage change in future spending		
	Linear	Quadratic	Piecewise
Delivery time (hr)	-0.081^{***}	0.014	-0.065^{***}
	(0.002)	(0.017)	(0.003)
Delivery time squared		-0.027^{***}	
		(0.005)	
Late time (hr)			-0.060^{***}
			(0.008)
Observations	91	91	91
R^2	0.939	0.955	0.962
Adjusted R^2	0.938	0.954	0.961
Residual Std. Error	0.009 (df = 89)	0.008 (df = 88)	0.007 (df = 88)

*p<0.1; **p<0.05; ***p<0.01

Note: Each observation summarizes the causal impact of delivery time on customer retention at minute level, from minute 60 to minute 150.

6 Alternative Explanations: Left-digit Bias and Plans

One alternative explanation is left-digit bias: customers pay more attention to the hour digit rather than the minute digit. If orders are promised to arrive before 5:00pm, then customers may perceive arriving at 4:49pm vs 4:59pm to be similar because the left-most digit remains the same, but perceive arriving at 4:59pm vs 5:09pm to be different because the left-most digits are different. This explanation is plausible if all orders are promised to arrive around the 60-minute mark.

Another explanation is that customers are disappointed because of disrupted plans. If customers have pre-determined plans and then experience late deliveries, customers may be disappointed because of a real disutility caused by disrupted plans rather than a psychological disutility caused by reference dependence. The disappointed customers may become less loyal to the platform or perceive the delivery time to be longer than they actually are, thus less likely to use the service in the future.[12] Because people tend to have plans in 30-minute or 60-minute blocks, this explanation implies that the impact of late deliveries will be much stronger if deliveries are promised to arrive around the 30-minute or 60-minute marks.

We show that these two explanations are not sufficient for explaining the data by leveraging variation in the minute digits of the promised delivery time. The digits of the promised arrival time is quite arbitrary in our dataset because the arrival minute depends on when exactly the order is placed. If a two-hour delivery order is placed at 13:43, it will be promised to arrive before 15:43. Let HH:MM be this promised arrival time. Figure 6 shows that the distribution of the minute-digit MM is almost uniformly distributed.

If left-digit bias or plans are the primary driving force of the asymmetric response, then the estimated loss aversion will be much stronger for orders that are promised to arrive around the 60-minute mark. We define an order to be around the 60$-$minute mark if the promised delivery time is 10 minutes before or after the 60-minute mark.[13] Figure 7 compares the impact of delivery time on retention based on whether the order is promised to arrive around the 60-minute mark. Even when the promised delivery time is not around the 60-minute mark, customers become disappointed when the delivery is late, suggesting that left-digit bias or plans are not sufficient for explaining the asymmetric response.

[12] This assumption already deviates from the utility specification in standard empirical models that assume current utility to be stable and independent of past dis-utility after controlling for belief. For the purpose of setting promises, distinguishing the sources of disappointment and loss aversion for the estimated parameter γ is not important because the structural model and the corresponding managerial implications is similar.

[13] Formally: MM$\in [0, 10] \cup [50, 60]$. For example, an order that is promised to arrive before 4:51pm falls into this category but an order promised to arrive before 4:49pm does not.

Figure 6: Distribution of minute-digit of the promised delivery time

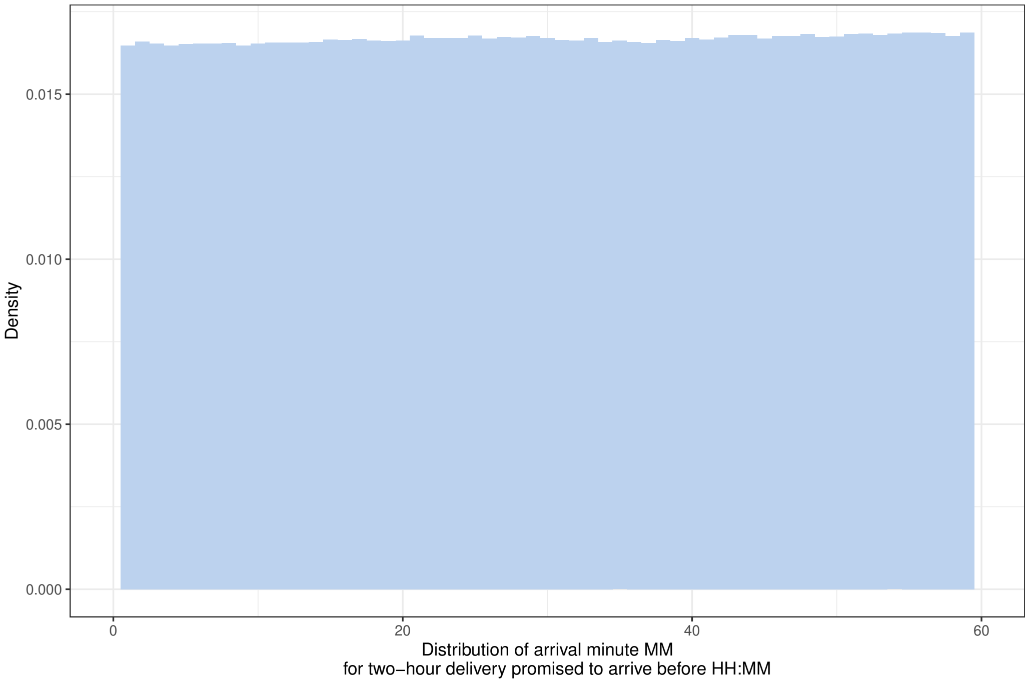
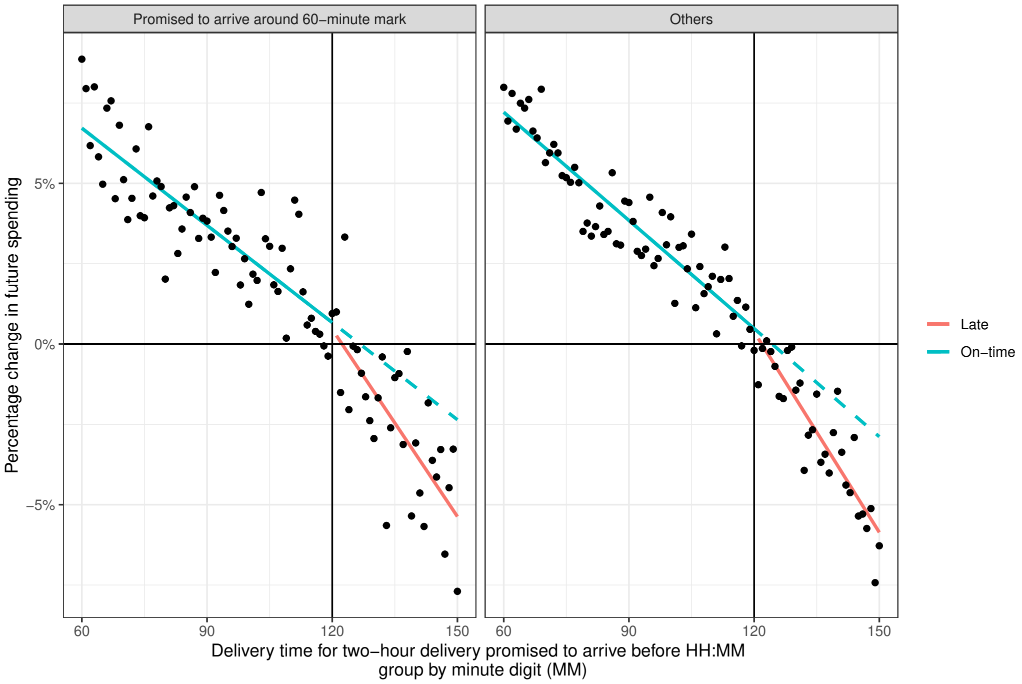


Figure 7: Impact of delivery time on retention, grouped by promised arrival minute



7 The Role of Customer Experience

Because customers have multiple purchase occasions, it is essential to construct a multi-period model for setting promises. A prerequisite to such a model is understanding how customers change their behaviors as they gain more experience. This section discusses how customer experiences affect the reference point r, the responsiveness β, and the loss aversion γ.

7.1 Evidence of promise-based reference points

As acknowledged by Kahneman and Tversky (1979), the reference point could be affected by customer expectation, which is often empirically approximated by average experience. In the context of service delivery, one possibility is that customers form their reference points based on the average delivery time they experienced in the past. For example, if customers on average receive their delivery in 90 minutes for orders promised to arrive within 120 minutes, then customers may expect the delivery to take around 90 minutes in the future. If these customers use the expected 90 minutes as reference points, they will be disappointed by an on-time delivery that takes 100 minutes.

To quantify the importance of promises versus experiences on the formation of reference points, we replicate our analysis in Section 5 based on customer experience. Recall that the analysis in Section 5 focuses on the main effect of delivery time on retention across all orders. Among these orders, some are placed when customers have no prior experience with the platform. Others are placed when customers already have sufficient experience with the platform. Some customers on average receive their delivery 30 minutes earlier than promises. Others receive their delivery right on time. Let n_i be the number of orders that customers have placed before placing the current order i, and let θ_i summarize the average deviation experienced by the customer in the past, grouped into 10-minute bucket.[14] We partition orders based on (n_i, θ_i) and use the partitioned data to estimate a conditional causal response curve:

$$\mu(D|n,\theta) = E[Y_i(D)|n_i = n, \theta_i = \theta]$$

This causal response curve allows us to answer counterfactual questions such as: for customers who have placed n orders and experience an average deviation of θ, how do their most recent delivery experiences affect retention? We use the generalized propensity score method to estimate the average potential retention $\mu(D|n,\theta)$ for all (D, n, θ), and evaluate how well a

[14] For example, all customers who experience an average delivery time of $[100, 110]$ minutes in the past, hence a deviation of $[10, 20]$ minutes, will be grouped into one bucket.

class of parametric heterogeneous piecewise linear models fit the estimated points:

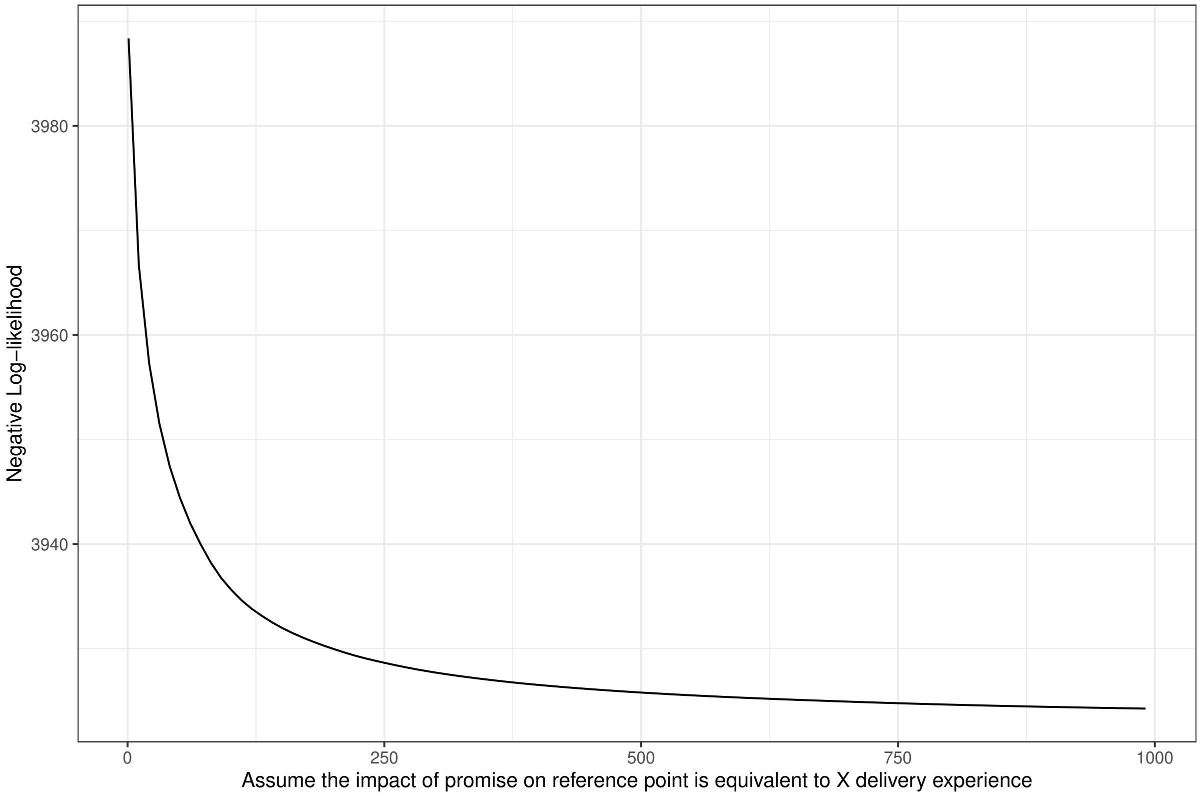
$$\mu(D|n,\theta) = \begin{cases} \alpha(n,\theta) + (D - r(n,\theta))\beta(n,\theta) & D \leq r \\ \alpha(n,\theta) + (D - r(n,\theta))[\beta(n,\theta) + \gamma(n,\theta)] & D > r \end{cases} \tag{7}$$

where we assume that the reference point is determined by:

$$r(n,\theta) = \frac{n}{n + w_{promise}}(Promise - \theta) + \frac{w_{promise}}{n + w_{promise}} Promise \tag{8}$$

$w_{promise}$ is an unknown parameter that summarizes the weight that customers place on promises relative to experiences. If $w_{promise}$ is infinite, the reference point only depends on promises and does not depend on past experiences. If $w_{promise} = 1$, a 2-hour promise is equivalent to experiencing one delivery that arrives around 2-hour. Figure 8 illustrates the log likelihood of the estimated response curve when assuming customers place different weights on promises. The model has better goodness-of-fit if the weight is extremely large, suggesting that past average experience has negligible impact on reference point.

Figure 8: Compared to promises, past average experiences have limited impact on reference points based on log-likelihood



Another way to demonstrate this promise-based reference point is to focus on a subset of experienced customers who used to receive delivery much earlier than promises, and test whether the reference point is still around promises. Appendix B demonstrates this method and shows that although these customers used to experience earlier-than-promised delivery times, they still use promises as reference points. For the remaining sections, we maintain this assumption that customers directly use promises as reference points, regardless of their past experience.

7.2 Evidence of learning

In addition to reference points, the baseline responsiveness to the most recent delivery time, β_i, may also depend on customer experience. To study how the the impact of delivery time is moderated by customer experience, we partition orders based on the customer experience measure n_i, and use the partitioned data to estimate a conditional causal response curve:

$$\mu(D|n) = E[Y_i(D)|n_i = n]$$

This causal response curve allows us to answer counterfactual questions such as: for customers who have placed n orders, how do their most recent delivery experiences affect retention? To summarize this effect, we parameterize this conditional response curve based on Equation 6:

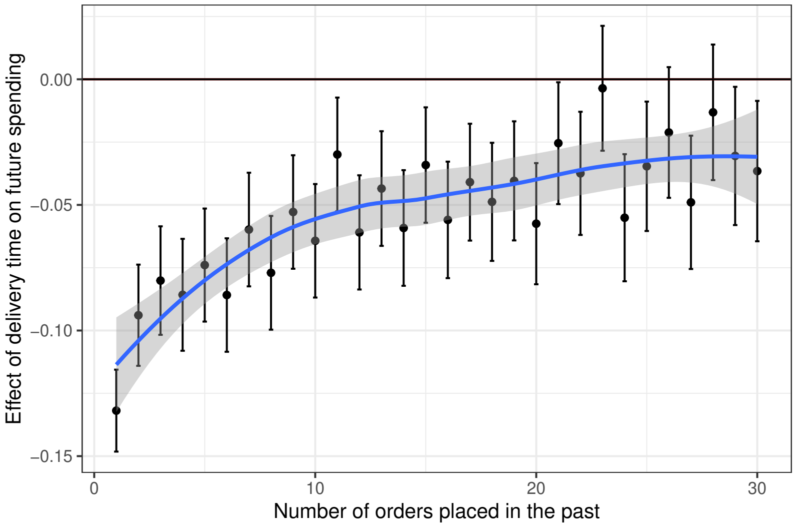
$$\mu(D|n) = \begin{cases} \alpha_n + \beta_n(D - r) & D \leq Promise \\ \alpha_n + (\beta_n + \gamma_n)(D - r) & D > Promise \end{cases} \tag{9}$$

where $\beta_n = E[\beta_i|n_i = n]$ summarizes impact of delivery time on retention when it is on-time for customers with experience n, and γ_n summarizes the loss aversion when the delivery becomes late.

Figure 9 summarizes how the base impact of delivery time on customer retention, β_n, is related to customer experience. For inexperienced customers, their most recent experience has a large impact on their retention. For the experienced customers, this impact is limited and closer to zero. This diminishing pattern is consistent with learning: more experienced customers are more likely to rely on their prior experience, thus less likely to update their belief based on one additional experience. For example, experiencing a 3-hour delivery may lead a new customer to expect future order to also take 3 hour, making the customers much less likely to reorder. In contrast, experiencing the same 3-hour delivery may not lead an experienced customer to expect future order to take 3 hours, if this experienced customer used to receive the delivery in 1-hour.

One limitation of Figure 9 is that it only provides suggestive evidence of learning. Appendix C discusses alternative explanations such as heterogeneity and selection, as well as possible ways to rule them out. For our structural model, we assume that customers use promise as a signal

Figure 9: Customer experience vs effect of delivery time on retention



and learn to interpret the signal as they gain more experience. Section 9.2 demonstrates the usefulness and the plausibility of this learning assumption by showing that such learning model can generate accurate counterfactual policy predictions.

7.3 Loss aversion vs experience

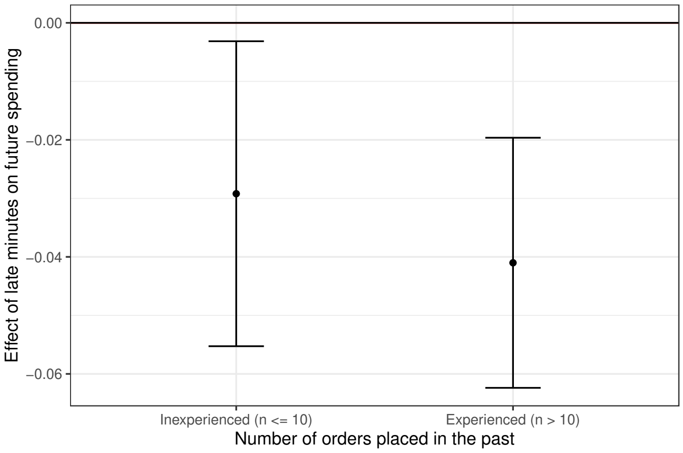
A related question is whether the degree of loss aversion γ_n depends on customer experience. We do not have enough data to accurately estimate the change of loss aversion over time.[15] Figure 10 compares the loss aversion for inexperienced customers who have placed no more than 10 orders against experienced customers who have placed more than 10 orders. We do not find evidence that experienced customers are less responsive. Therefore, we assume that the degree of loss aversion does not depend on experience for our subsequent structural model.

8 Model

We first discuss a general framework to highlight the key mechanisms through which promises affect customer decisions. Then we discuss a parsimonious and tractable model to inform firms on designing promises.

[15] Identification of loss aversion relies on the variation in late delivery minutes. Because the majority of deliveries in our dataset arrives on time, such variation in late minutes is limited after partitioning based on customer experience.

Figure 10: Customer experience vs effect of late minutes time on retention



8.1 General framework

Consider a customer i who has a certain prior belief about the distribution of the service quality D_{it} at time t:

$$D_{it} \sim F(Promise_{it}; \Theta_{it}).$$

where $Promise_{it}$ is the promise presented to the customer, and F is an arbitrary distribution parameterized by promises and prior belief Θ_{it}. The prior belief Θ_{it} may include how the customer interprets and uses promises. For example, a promise of arrival within 2 hours can be interpreted as having a 90%, 95%, or 100% probability of arriving within 2 hours.

This belief affects the customer's current purchase decision through changing the expected utility of using the service, $\bar{U}_{it}$:

$$\bar{U}_{it} = g(F(Promise_{it}; \Theta_{it})) + Sat_{i,t-1} + \epsilon_{it} \tag{10}$$

where g is a function that summarizes how the distribution of quality, or certain moments of the distribution, affects the expected utility. For example, the customer may prefer faster average delivery time. The customer may also prefer the delivery time to have smaller variance so it is easier to make plans.

Additionally, customer past cumulative satisfaction or goodwill, $Sat_{i,t-1}$, may directly affect the utility of using the service. Intuitively, even if experienced customers have stable belief Θ_{it} over the delivery time, their preferences for using the service may still decrease significantly if they experience a late delivery in an important occasion.

Equation 10 is a general specification that captures two key mechanisms through which current promises affects future purchases: belief and satisfaction. First, after experiencing the actual quality D_{it}, customers may learn how to interpret promises based on how the promised quality is different from the actual quality:

$$\Theta_{i,t+1} = Learning(\Theta_{i,t}, Promise_{it}, D_{it})$$

which in turn affect the belief in the future distribution, $F(Promise_{i,t+1}; \Theta_{i,t+1})$.

Second, promises may directly affect current satisfaction and hence future utility of using the service:

$$Sat_{it} = s(Promise_{it}, D_{it}, Sat_{i,t-1}).$$

where s is a function that characterizes how the cumulative satisfaction evolve as customers experience new promised and actual quality. To summarize, the impact of promises on future expected utility can be structurally decomposed into 1) an indirect effect on future belief and 2) a direct effect on future utility:

$$\frac{\partial \bar{U}_{i,t+1}}{\partial Promise_{it}} = \underbrace{\frac{\partial g(F(Promise_{i,t+1}; \Theta_{i,t+1}))}{\partial \Theta_{i,t+1}}}_{\substack{\text{Effect of belief} \\ \text{on expected utility}}} \underbrace{\frac{\partial \Theta_{i,t+1}}{\partial Promise_{it}}}_{\substack{\text{Effect of promises} \\ \text{on future belief}}} + \underbrace{\frac{\partial Sat_{it}}{\partial Promise_{it}}}_{\substack{\text{Direct effect of promises} \\ \text{on future utility}}}$$

Worth noting is our general framework does no impose assumptions on how the expected utility is formed. A common alternative specification is to assume that customers have exogenous utility of receiving a certain quality of service $u(D)$, and the expected utility $E[u(D)|F]$ can be derived based on the distribution of quality F. Although our framework can account for this specification, this specification is implausible in the context of delivery time, because the utility of potentially receiving a delivery exactly at a *certain* minute is endogenous and depends on customers' plans.[16] Due to this complication, we abstract away from how the expected utility is formed. This abstraction is sufficient for the purpose of designing promises because customer purchase decisions are based only on the expected utility and how it is affected by the belief in F.

8.2 Parsimonious model

Next, we discuss a parsimonious and tractable model to help inform how firms should design promises. For the actual and promised quality, we focus on the log actual delivery time D_{it} and log promised delivery time $Promise_{it}$. Using this log transformation allows us to account for the constraint that delivery time must be positive. It is also consistent with the asymmetrical distribution in Figure 3.[17]

[16] To formally model customer plan, we can allow customer utility to depend on their plans as well as the realized quality $u(D, plan)$. Customers will choose plans optimally based on the distribution of delivery time, such that the expected utility is $g(F) = E[u(D, plan^*(F))|F]$.

[17] An alternative specification is to assume that the delivery time itself is normally distributed. This is implausible because it implies that delivery time could be negative and the distribution would be symmetrical.

Assume customer i believes the log delivery time D_{it} is drawn from a normal distribution:

$$D_{it} \sim N(Promise_{it} - \theta, \sigma_{i,\epsilon}^2)$$

where $Promise_{it}$ is the promised log delivery time, and θ summarizes how, on average, the actual delivery time may be different from the promised delivery time. This assumption implies that customers do pay attention to promised delivery time, which is supported by our empirical finding. This assumption also implies that promises are informative of the actual delivery time, which is also true for most firms.[18] For example, $\theta = 0.1$ implies that the customer believes the average delivery time is approximately 10% faster than promises. Customers have uncertainty over this deviation and have a prior belief that the deviation follows:

$$\theta \sim N(\theta_{it}, \sigma_{it}^2).$$

Let $I_{it} = \{\theta_{it}, \sigma_{it}^2, \sigma_{i,\epsilon}^2, Promise_{it}\}$ be the information set available to the customers at time t. Assume the expected utility of using the service depends on the average and variance of the delivery time, such that:

$$\bar{U}_{it} = \alpha_i + \beta_i^{time} E[D_{it}|I_{it}] + r_i Var[D_{it}|I_{it}] + \delta X_{it} + Sat_{i,t-1} + \epsilon_{it}$$

where α_i is the individual-specific demand, β_i is the time sensitivity to delivery time, r_i summarizes risk aversion, X_{it} summarizes customer characteristics at the time when they place the order that may also affect the demand for the service, and ϵ_{it} is the demand shock that is observed by consumers but not observed by econometricians. Customers will make purchase decisions if the expected utility is above 0

$$Y_{it} = \begin{cases} 1 & \bar{U}_{it} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

After experiencing the delivery time, a standard Bayesian learning model would predict that customers will update their belief based on Bayes rules:

$$\theta_{i,t+1} = (1 - \lambda_{it})\theta_{it} + \lambda_{it}(Promise_{it} - D_{it})$$

where $\lambda_{it} = \frac{1/\sigma_{i,\epsilon}^2}{1/\sigma_{i,\epsilon}^2 + 1/\sigma_{it}^2}$ summarizes how much weight Bayesian customers would place over their most recent experience when updating their belief. This updating rule is symmetrical and linear, and cannot rationalize our asymmetric pattern around promises. The uncertainty $\sigma_{i,t+1}$

[18] The model may be implausible in a counterfactual scenario or equilibrium where promises are completely uninformative, such that customers do not use promises to form their belief.

will also be updated smoothly $\frac{1}{\sigma^2_{i,t+1}} = \frac{1}{\sigma^2_{it}} + \frac{1}{\sigma^2_{i,\epsilon}}$, and cannot rationalize empirical finding that the response curve is kinked around promises.

One way to rationalize the pattern is to assume that for customers who have recently made a purchase $Y_{it} = 1$, their satisfaction is reference-dependent and asymmetrically depends on whether the delivery arrives earlier than the promise:

$$Sat_{it} = \begin{cases} \beta^{gain}(D_{it} - Promise_{it}) & D_{it} \leq Promise_{it} \\ (\beta^{gain} + \gamma)(D_{it} - Promise_{it}) & D_{it} > Promise_{it} \end{cases} \tag{12}$$

For customers who have not recently made a purchase $Y_{it} = 0$, we assume their satisfaction stays the same as the previous period, such that $Sat_{it} = Sat_{i,t-1}$.[19]

8.3 Model implications

This simple model of learning with reference dependence can rationalize our empirical findings that the impact of the most recent delivery time on expected utility is kinked around promises:

$$\frac{\partial \bar{U}_{i,t+1}}{\partial D_{it}} = \begin{cases} \beta^{time}\lambda_{it} + \beta^{gain} & D_{it} \leq P_{it} \\ \beta^{time}\lambda_{it} + \beta^{gain} + \gamma & D_{it} > P_{it} \end{cases} \tag{13}$$

Because the weight λ_{it} decreases as customers gain more experience, this learning specification also rationalizes our finding that the impact of the most recent customer experience on future purchase diminishes as customers gain more experience:

$$\begin{aligned} \frac{\lambda_{i,t+1}}{\lambda_{i,t}} &= \frac{\sigma^2_{i,t+1}}{\sigma^2_{it}} = \frac{1}{\sigma^2_{it}(\frac{1}{\sigma^2_{it}} + \frac{1}{\sigma^2_{i,\epsilon}})} \\ &= \frac{1}{1 + \frac{\sigma^2_{it}}{\sigma^2_{i,\epsilon}}} \\ &< 1 \end{aligned} \tag{14}$$

The model also predicts that the impact of late arrival will not go to 0 as customers become more experienced because of the loss aversion parameter γ, which is supported by Figure 10.

A related question is whether our empirical finding can be rationalized by standard Bayesian learning model with non-linear utility specification, without imposing this additional reference-dependent satisfaction term. For example, customers may have quadratic preferences with respect to delivery time such that $U_{it} = \alpha_i + \beta_{i1}D_{it} + \beta_{i2}D^2_{it} + \epsilon_{it}$. Under this quadratic specification, the impact of delivery time on future expected utility will be a smooth quadratic

[19] This assumption is made for tractability and ease of estimation. Future iteration of the model will consider the satisfaction to be a cumulative function of past experience with an estimable discount factor.

curve as well, which is rejected by Figure 5 that supports a response curve that is kinked around promises. Appendix D.1 illustrates that even if customers have plans such that the delivery time has a discontinuous and asymmetric effect on customers' *current* utility, its impact on *future* expected utility will not be kinked around promises due to uncertainty. Therefore, this kinked response curves cannot be easily rationalized by standard Bayesian learning with asymmetric utility specifications.[20]

8.4 Identification without variation in promises

A benefit of imposing the learning structure is that variation in promises is not required for estimating the model, and the estimated model can be used to simulate counterfactuals on what happens when promises change. This is because the model assumes that promises affect customers' current decisions through changing their belief about the average delivery time. Therefore, we can estimate the impact of changing promises if we can estimate how the belief in the average delivery time affect purchase decisions, β^{time}. This parameter can be identified if there is variation in the belief over the average delivery time, $E[D_{it}|I_{it}]$. This variation in $E[D_{it}|I_{it}]$ can not only be driven by variation in promises, but can also be driven by variation in past experienced delivery time, which we directly observe in the data.

Therefore, variation in promises can either be used to improve estimation efficiency or be used to validate the model. We do have access to an A/B test that randomizes promises presented to customers. Given that one concern of using structural model for empirical market design is that they may have limited credibility due to simplified parametric assumptions, we decide to use it for validation rather than estimation in Section 9.2 to demonstrate the credibility and value of our structural model.

8.5 Estimation

One estimation challenge is heterogeneity of purchase patterns across time. For example, urban customers who start to use the service during the pandemic have different usage patterns from suburban customers who start to use the service before the pandemic. To account for this user-time heterogeneity, we define customer cohorts based on the locations and timing of customers' initial orders, and use the average purchase probability at the cohort-market-week level as a control.[21] Intuitively, customers from the same cohort should have similar usage patterns over

[20] Another way to rationalize the kinked pattern is to deviate from standard Bayesian learning and assume that customer update their belief asymmetrically around promises. We discuss this alternative specification and testable implications in Appendix D.2

[21] For linear regression, this control can be interpreted as cohort-market-week interacted fixed effect.

time. Their future purchase behaviors may differ because they experience different delivery times. We focus on a subset of large cohorts that have at least 100 customers and have more than 4 months of observations in the sample. We obtain a total of $77,778$ customers from 464 different cohorts. To further account for unobserved user heterogeneity, we assume that the individual preference is drawn from $\alpha_i \sim N(\alpha_0, \sigma_\alpha^2)$.

Another estimation challenge is that delivery time is not the only service attribute that customers care about. As customers gain more experience, customers will become familiar with other dimensions of service attributes unobserved by researchers. For example, if a customer becomes more experienced with the mobile app or websites, the customer can search the items they want faster either due to increased familiarity or better personalization, and hence more likely to use the service. To address this issue, we control for the number of orders customers have placed using a parametric function of customer experience $Familiarity(n) = -\frac{1}{n+1}$.[22] This control captures how customer familiarity and preference for other service attributes evolves as they gain more experience.[23]

The existence of other unobserved service attributes makes it difficult to identify the prior belief about delivery time: even if we observe that a customer uses the service more frequently as the customer gains more experience, we cannot distinguish whether it is driven by 1) low prior belief about delivery time or 2) low prior belief about other service attributes. However, this identification issue does not affect the counterfactual simulation because the model implies that the marginal impact of promises on utility does not depend on this prior θ_{i0}. Additionally, because we directly observe delivery time, we can still identify the prior uncertainty. Following Sriram et al. (2015) we infer the precision $\sigma_{\epsilon,i}$ based on the actual variation of delivery time,[24] with the assumption that customers have some rational expectation about the variation in delivery time. The variation in delivery time also allows us to identify the prior uncertainty, σ_0^2, based on how customer responsiveness to delivery time diminishes as they gain more experience.[25]

Table 2 presents the key coefficients. The time sensitivity parameter $\beta = -0.252$ is negative, supporting that customers prefer faster delivery time. The risk aversion parameter $r = -0.506$,

[22] We've experimented with different parametric specifications and find that $Familiarity(n) = -\frac{1}{n+1}$ has the best in-sample goodness-of-fit.

[23] This function has a causal interpretation if the unobserved customer heterogeneity and auto-correlated demand shocks are sufficiently accounted for using cohort-market-week fixed effect and random coefficients.

[24] In Sriram et al. (2015), the precision is inferred based on the actual variation of service quality at household level, which works if customers are observed repeatedly. Because in our sample some customers are only observed once, we use the predicted variation in Section A, which works even if a customer is observed in the sample once.

[25] Intuitively, if the impact of delivery time has a large effect on repurchase for first-time customers but significantly smaller impact for second-time customers, then it suggest the prior uncertainty must be relatively large.

suggesting that customers dislike uncertainty. $\beta^{gain} = 0.024$ is small, suggesting customer satisfaction does not improve much as they experience early arrival. Therefore, the benefit of arriving early mainly improves customer repurchase through learning and belief, rather than through satisfaction. The loss aversion parameter $\gamma = -0.316$ is negative, suggesting that late arrival has a significant negative effect on futuere repurchase. The prior uncertainty is $log(\sigma_0^2)$ is positive, suggesting that customers initially have high uncertainty over the average delivery time.

Table 2: Estimation results for learning model with reference dependence

	Dependent variable: Purchase	
	Estimate	Std. Error
Time preference		
Intercept (α)	3.865	(0.662)
Time sensitivity (β)	-0.252	(0.036)
Risk aversion (r)	-0.506	(0.141)
Log prior uncertainty ($log(\sigma_0^2)$)	3.681	(2.818)
Unobserved heteorgeneity ($log(\sigma_\alpha^2)$)	0.263	(0.006)
Reference-dependent satisfaction		
Response to gain (β^{gain})	0.024	(0.025)
Loss aversion (γ)	-0.316	(0.100)
Covariates		
Familiarity ($-\frac{1}{n+1}$)	2.981	(0.264)
Cohort-week-control	11.774	(0.118)
Log likelihood	-282732.8	

9 Validation

We leverage a policy change and an A/B test to validate our models, testing whether our model can successfully predict what is going to happen when the promise policy changes. This exercise is particularly meaningful for demonstrating how a structural economic model complements a prediction-based machine learning model because the data used for estimation does not have any variation in the promised delivery time. Due to this lack of variation, a pure prediction-based model cannot answer counterfactual questions related to changing the promised delivery time. However, because the structural model allows us to identify preference for time β and loss aversion γ, we can predict the impact of promises based on how promises affects customer belief in the delivery time.

9.1 Validation data 1: Policy change

We observe a policy update that changes the delivery options available to the customers when they check out. Before the policy change, the platform first calculated the ETA of the delivery based on the order characteristics and then rounded the ETA to either 1-hour, 2-hour, or 5-hour. After the policy change, the platform removed the rounding and provided a more precise ETA. For example, if the ETA for an order was 1:45, the earliest available delivery option was 2 hours before the policy change, but 1:45 after the policy change.

The policy change allows us to validate whether customers still use 2 hour as reference points when the promised delivery time is no longer 2 hour. Our model predicts that when the promise is no longer 2-hour, 1) customers will not change their responsiveness to delivery time around 2-hour, and 2) customers will change their responsiveness around promises.

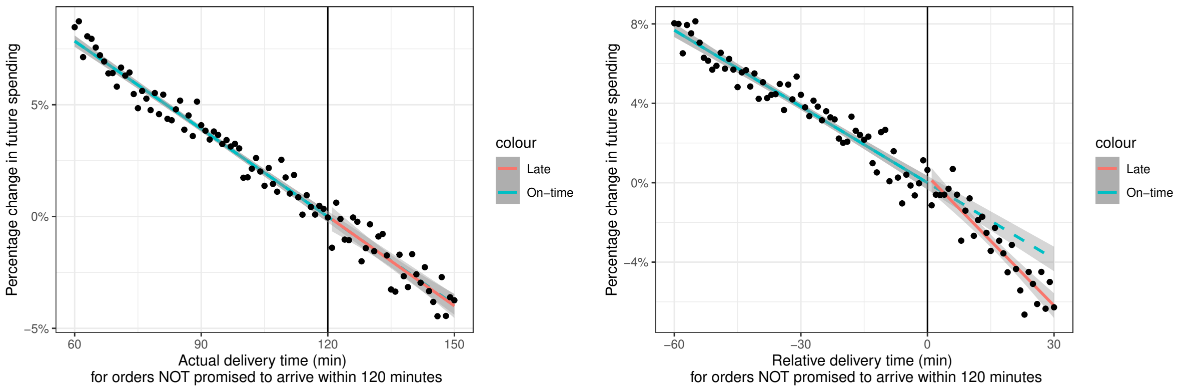
To validate that customers do no change their responsiveness around 2 hours, we estimate causal response curve using the orders placed by customers after the policy change. We focus on the subset of orders of which the promised delivery times are different from 2 hours, and apply the estimation technique discussed in Section 5 on this subset of orders. Figure 11a demonstrates the impact of deviating from 2 hour for this subset of orders, and shows that customers do not suddenly become more responsive when the delivery takes longer than 2 hour.

To validate that customers change the responsiveness around promises, we use the same subset of orders but focus on estimating the impact of relative delivery time, rather than absolute delivery time, on customer retention

$$\mu(D - Promise) = E[Y_i(D - Promise)]$$

The estimation is similar to the exercise in Section 5, except that the continuous treatment variable of interest is the relative delivery time rather than the absolute delivery time.[26] Figure 11b demonstrates the estimation result, which shows that customers become more responsive when delivery arrives later than promises.

Figure 11: Impact of absolute vs relative delivery time on retention when promised delivery time is no longer 2-hr



(a) Absolute delivery time

(b) Relative delivery time

9.2 Validation data 2: A/B test

One limitation of the previous validation exercise is that it still relies on Assumption 1 that the delivery time is independent of potential customer retention after conditioning on observables. The analysis also focuses on the long-term intensive margin of customer retention conditional on making purchases, and does not validate the short-term extensive margin on how promises affect current purchases. This section leverages an A/B test conducted during the policy change to directly validate the impact of promises on immediate customer purchase decisions. Such A/B tests are often conducted by E-commerce platforms to improve customer experience and firm revenue.

The A/B test changes the delivery options available to customers when customers are about to checkout. The A/B test randomly assigns customer-visit into 5 groups. For customers in the control group, their promised delivery time is calculated based on an existing algorithm. There are four treatment groups. Customers in these treatment groups respectively receive promised delivery times that are 20-minute earlier, 20-minute later, 40-minute later, and 60-minute later than the control group. This A/B test was intended to measure the short-term impact of promises on current customer purchases.

[26] When there is no variation in the promised delivery time, it is not necessary to distinguishing these two treatments because they are perfectly correlated.

Figure 12 illustrates the result of changing promises, suggesting that customers are more likely to use the service if they see faster promised delivery time. This finding is consistent with our model estimates that customers prefer faster delivery time: $\beta < 0$.

Figure 12: Impact of changing promises on current purchases

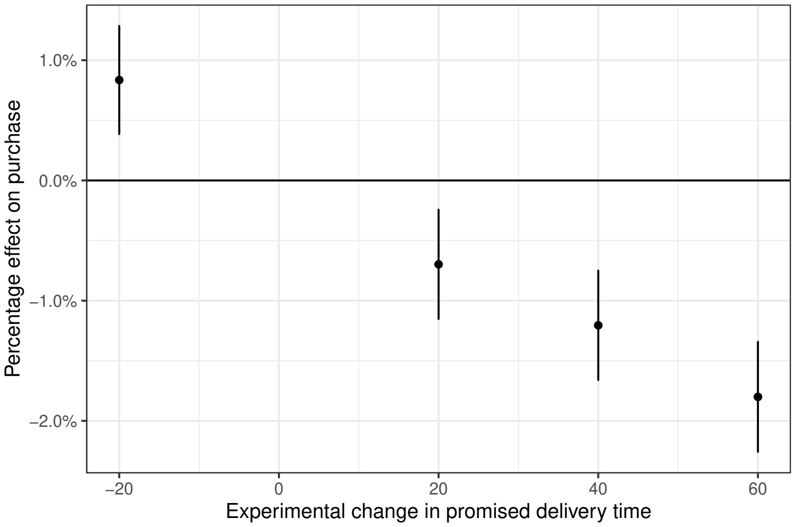


Table 3 shows that our model not only correctly predicts *whether* customers prefer faster deliveries, but also the *degree* to which they prefer faster deliveries. The experimental result shows that an hour increase in the promised delivery time would decrease the immediate purchase by 1.91% for customers who are about to check out. We ask whether our structural model of learning can make similar predictions. Based on the purchase probability of the control group,[27] and the time sensitivity β estimated by our model, our model generates a similar estimate, predicting that an hour increase in the promised delivery time would decrease the purchase probability by 2.13%.

9.2.1 Experiment limitation

A related question is why a model of how customers internalize firm promises is needed if firms can run such an experiment that randomizes promises. Although this type of experiment is valuable for measuring the immediate impact of promised delivery time on purchases, it has limitations in measuring the long-term impact of promises on customer retention and satisfaction because of institutional limitation: the orders are fulfilled based on a queuing system that

[27] Because an average customer is different from a customer who is about to check out, the validation exercise needs to account for this heterogeneity and selection. We assume that these customers have same time sensitivity β, but different baseline purchase probability α, where α can be estimated based on the average purchase probability of the control group.

Table 3: Model forecast vs experimental results

	Effect of promises on current purchases	
	Model forecast	Experiment estimate
Promise (hr)	-2.13%	-1.91%
95% Confidence interval	$(-2.70\%, -1.55\%)$	$(-2.20\%, -1.62\%)$

Note: The forecast is obtained based on the estimated time preference β and the checkout probability of the control group.

is affected by promises, and customers with faster promises will be prioritized. Therefore, randomly giving faster promises to treated customers will decrease the delivery time for the treated customers and increase the delivery time for the control customers. Because customer retention could be affected by the actual delivery time, the experiment violates the SUTVA condition, and cannot be used to correctly forecast the long-term impact of changing promises.

9.3 Discussion: value of structural model vs predictive machine learning

Our validation procedure helps illustrate the value of structural model. One common concern of structural model is that many parametric assumptions are imposed to ensure tractability. These restrictive assumptions imply parametric structural models may have lower goodness-of-fit than non-parametric predictive machine learning models. However, the value of the structural model partly depends on its ability to simulate counterfactual policies: even if there is no policy variation in the existing training data, the structural model can forecast what would happen when the policy changes, but a purely predictive model cannot. Our validation exercise helps illustrate this point: when there is no variation in promises in the training/estimation data (all promises are “within 2 hour”), a predictive model cannot forecast what would happen when promises change. However, by leveraging economic and psychological theory - that promises and past experience affect purchase decisions through belief and customers are reference-dependent, we can build a structural model that successfully forecasts what would happen when the promise policy changes.

Therefore, machine learning and structural models are complementary in our exercise. We use the machine learning models to estimate the generalized propensity scores, which help generate non-parametric empirical evidence that is consistent with reference dependence and learning. Then we build a structural model based on these findings, allowing us to simulate counterfactual to inform policy designs.

10 Counterfactual

This section simulates counterfactual promise policies: what would happen if promises are made such that $X\%$ of deliveries arrive on-time. We consider a fixed set[28] of customers who start to consider using the grocery delivery service in week 0, and examine how their purchase decisions evolve over time under different promises. Figure 13 demonstrates how the weekly purchase probability change over time if promises are made such that 10%, 50%, or 90% of the deliveries arrive on time. To illustrate the difference in purchase probability, the right figure summarizes the relative purchase probability using the promise with 50% on-time probability as a benchmark.

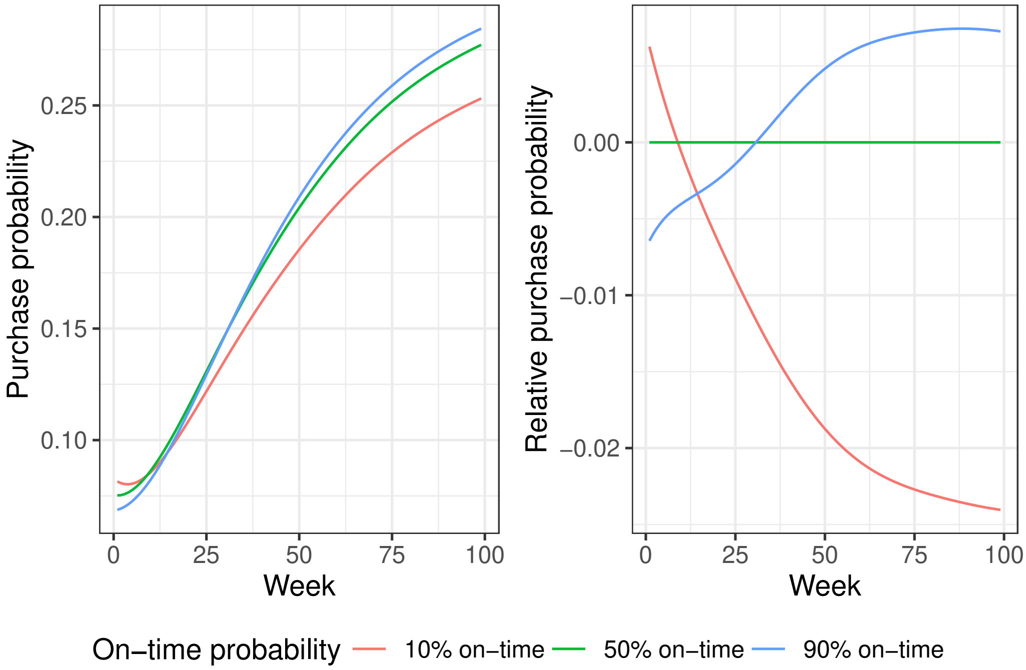


Figure 13: Relative purchase over time under different promises

In all cases, as customers become more experienced, they are more willing to use the service because they have less uncertainty over the delivery time, and they are more familiar with the service. However, although an aggressive promise with 10% on-time probability is more likely to attract customers in the first few weeks, this aggressive promise leads to fewer repurchases

[28] We consider a fixed set of customers because it is not costless to get customers to consider a product. Firms may spend money advertising and nudging customers to download their mobile apps and registering the accounts. It is therefore valuable to examine how often these customers make purchases.

in the long run. In comparison, although a conservative promise with 90% on-time probability does not immediately attract many customers, it leads to more repurchases in the long run. Figure 14 summarizes the relative cumulative purchases over a 100-week window for different promise strategies. The cumulative purchases are maximized if the promises are made such that 82% of deliveries arrive on-time.

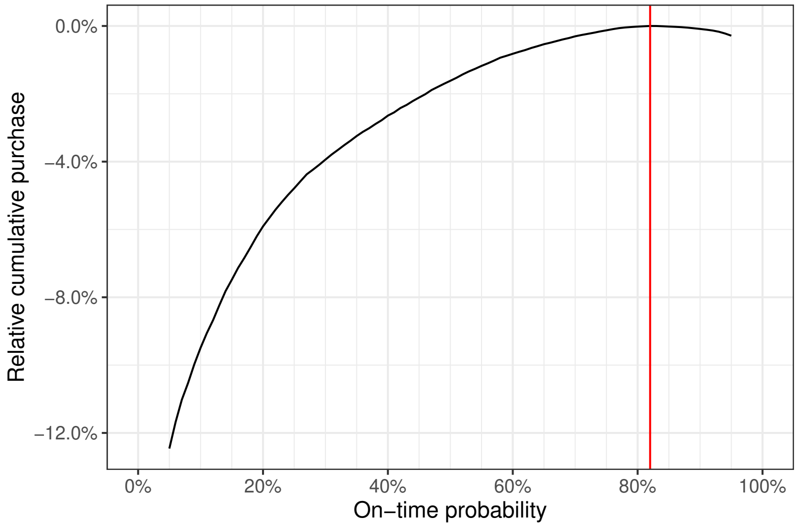


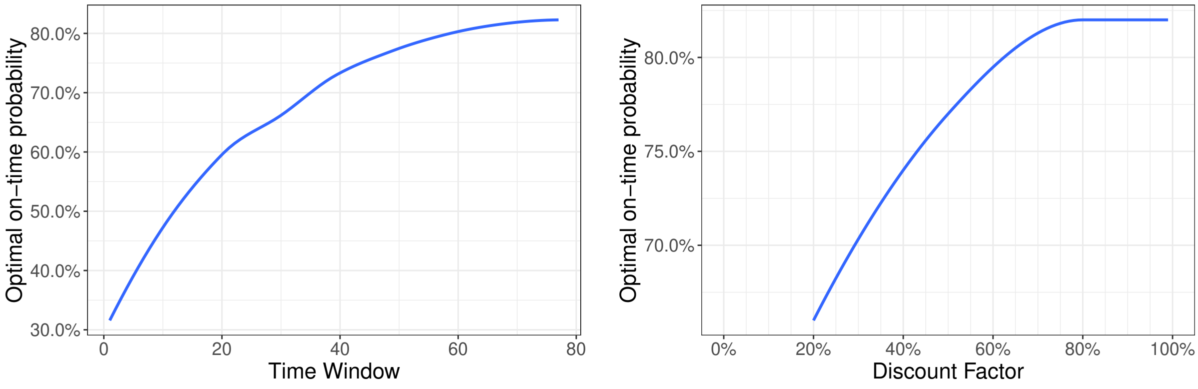
Figure 14: Relative cumulative purchases over 100 weeks

The optimal promise strategy depends on the window of analysis, which is relevant because some customers may only need online grocery service for a few months.[29] Figure 15a illustrates when the window of analysis is narrower, the optimal promise is more aggressive. This is because the long-term benefit of giving conservative promises is relatively small.

Because firms may care about the cumulative discounted revenue $\sum_{t=0}^{T} DiscountFactor^t Y_t$, the discount factor also affects the optimal strategy. This discount factor is usually assumed to be above 90%, meaning that \$100 a year later is worth at least \$90 today. Figure 15b compares the optimal promises under different annual discount factors. When the discount factor is above 80%, the optimal on-time probability is above 80%, suggesting that given reasonable discount factor, the long-term benefit of giving conservative promises to increase satisfaction is large even if it is discounted.

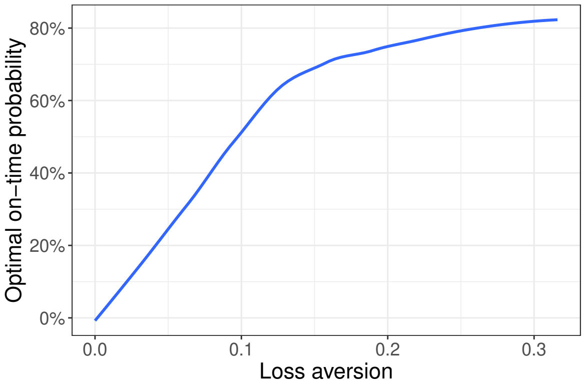
To highlight the importance of correctly estimating loss aversion, we demonstrate how the optimal promise may be different if firms incorrectly assume different values of loss aversion. Figure 15c shows that if firms under-estimate the degree of loss aversion, they may make promises that are too aggressive.

[29] For example, some customers may only need grocery service during a few COVID months.



(a) Time window vs optimal promise

(b) Discount factor vs optimal promise



(c) Optimal promise given different loss aversion

10.1 Expectation-based reference points

Because our empirical findings support that reference points are around promises, our structural model assumes a promise-based reference point. An alternative specification is to assume that customers form their reference points based on expectation, which could be affected by past average delivery time experienced by customers. Although our empirical findings do not support such expectation-based reference points, we illustrate what would happen if firms make such assumptions. Figure 16 summarizes how the purchase probability, and the difference in purchase probability, evolves over time. In all cases, customers are also more likely to reuse the service as they become more experienced. However, compared to promise-based reference points in Figure 13, the difference in purchase probability is much smaller in the expectation-based reference points, ranging from -0.6% to 0.6%. This is because customers will eventually learn what the average actual delivery time is, and use this average delivery time as reference points, regardless of what promises are given to customers initially.

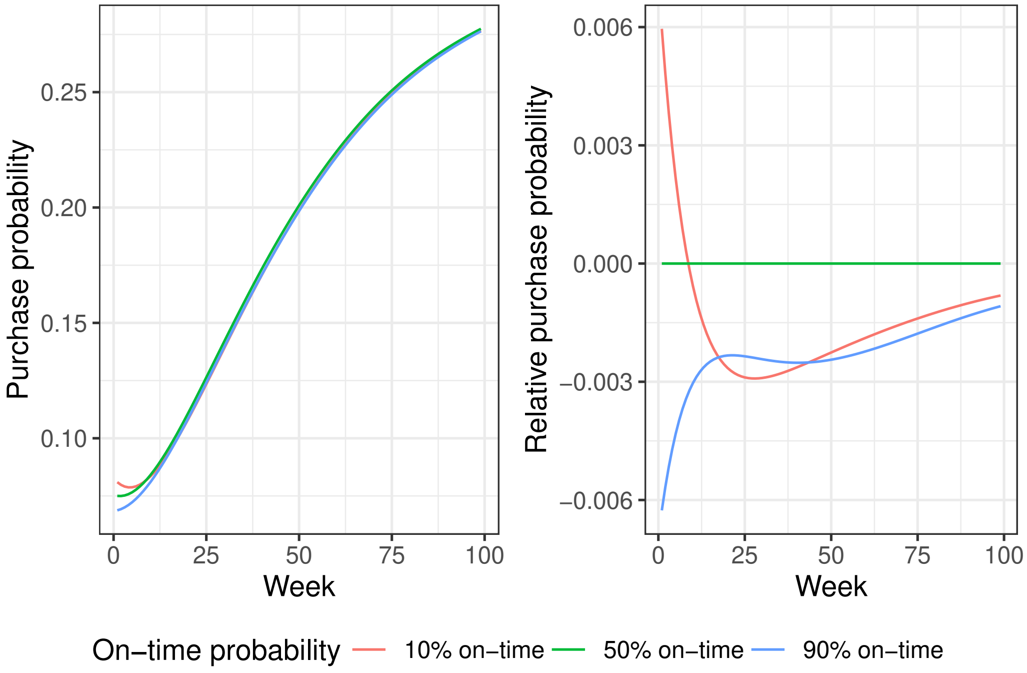


Figure 16: Relative purchase over time given expectation-based reference points

Figure 17 illustrates how the cumulative purchases are different given different promise strategies. The cumulative purchases are maximized when promises are made such that 64% of orders arrive on time. More conservative promises will lead to fewer cumulative purchases because the benefit of conservative promises on customer satisfaction is limited: customers will eventually learn the average delivery time, and will become disappointed when the delivery arrives later than this average delivery time, even if the delivery arrives on time.

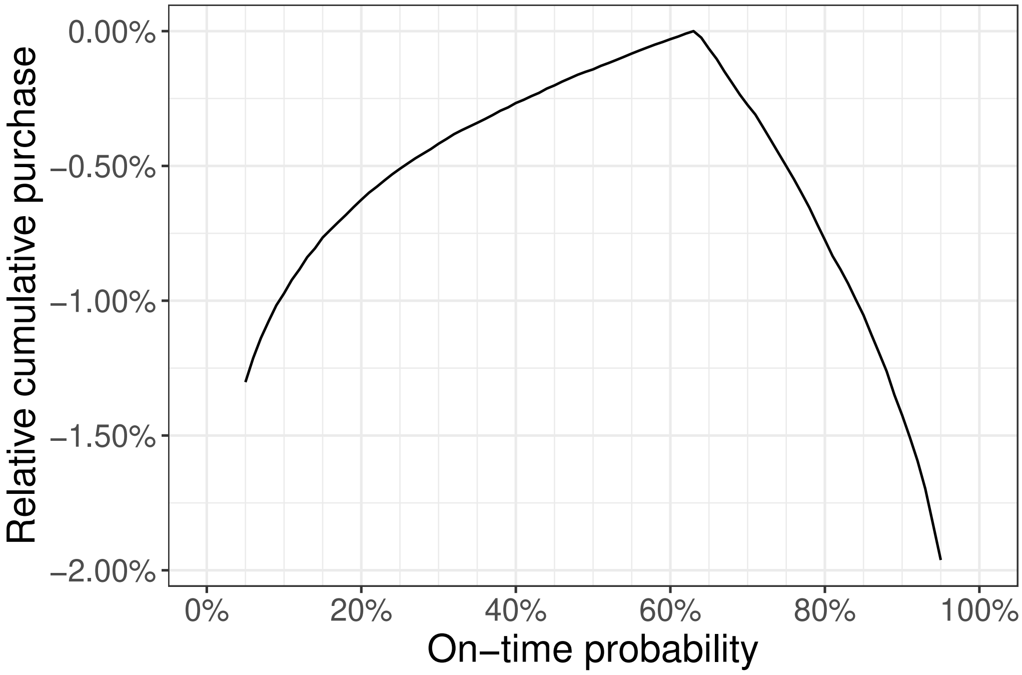


Figure 17: Relative cumulative purchases given expectation-based reference points

10.2 Revenue implication

Assume that purchases can be used to approximate revenue and the platform wants to maximize the total revenue, Table 4 summarizes the optimal promise strategies, and what happens when the platform misspecifies loss aversion or reference point. Given the correct parameters, the revenue is maximized when 82% of deliveries arrive on time. When firms under-estimate loss aversion or when firms incorrectly assume expectation-based reference points, they make promises that are too aggressive, leading to a relative revenue decrease of 0.60% to 4.24%. This is economically meaningful for firms because improving promises is almost costless: we consider holding the distribution of actual delivery time constant, such that the fulfillment cost remains similar. To understand the economic magnitude of this improvement, while Instacart itself has not publicly disclosed revenue rates, by some third-party estimates,[30] the platform reportedly earned 1.8 billion dollars in revenue in 2021. Therefore, if the platform underestimates loss aversion or mis-specify reference points in designing promises, the suboptimal promise strategy could translate to 10.69 to 76.27 million dollars loss in annual revenue.

[30] https://www.businessofapps.com/data/instacart-statistics/

Table 4: Optimal promise strategies with different assumptions

	Promise	Revenue Difference	
	(On-time probability)	Percent (%)	Dollar($)
Optimal	82%		
Underestimate loss aversion			
Underestimate by 80%	28%	−4.24%	−76.27 million
Promise median delivery time			
	50%	−1.62%	−29.09 million
Misspecify reference point			
Expectation-based reference	64%	−0.60%	−10.69 million

11 Conclusions

Our paper presents an empirical framework that incorporates prospect theory into designing promises. A fundamental challenge of applying prospect theory for market design is that the reference point is not directly observed by researchers: it is unclear what is perceived as a loss (Barberis (2013)). We present a novel technique to identify reference points by leveraging the property that customer response is kinked around reference points due to loss aversion. We find that customers use promises as reference points, and are 92% more responsive when the delivery arrives later than promises.

Motivated by our empirical findings, we construct a model of learning with reference dependence for designing promises. Our model is capable of simulating counterfactuals under different promise strategies. We leverage a policy and an experiment to show that even though our model is estimated based on data that has no variation in promises, it is capable of predicting how customers would respond when promises change. The validated model allows us to derive optimal promise strategies given different constraints. We illustrate that identifying the reference points and the corresponding loss aversion is crucial for designing promises. If firms incorrectly assume that customers form the reference point based on their average past experience or customers are not loss-averse, firms may set promises that are too aggressive, hurting customer retention and leaving millions of dollars on the table.

Our model is not without limitations. We consider a setting where the inherent variability of a service is large: neither firms nor customers can perfectly predict the actual delivery time. This inherent variability may affect reference formation. Without this variability, customers may be more likely to form reference points based on their past average experience. For example, a customer who always receive their delivery *exactly* around 90 minutes may use 90 minutes as a reference point even if the promised delivery time is 2 hours. This inherent variability is also required for our framework to identify reference points. Therefore, our framework may not be applied in consumer packaged goods where the variation in product quality is either small or unobserved. However, our framework can be still applied in industries such as E-commerce, ride hailing, and airlines, where the quality of the product or service often varies.

For designing promises, our model considers a simple setting where customers are making a binary choice of whether to purchase given one promise. As illustrated by our validation exercise, this binary specification is sufficiently useful in predicting what happens when promises change and illustrating the key trade-offs between customer acquisition and customer retention. In practice, customers are often given a menu of promises, and customers can also choose another competitor who may also strategically set promises. Our empirical findings on how customers respond to deviation from promises are still essential for designing promises in these settings. Future works hope to extend the existing framework to account for the design of promise menus and competition.

A Implementation of generalized propensity score method

To implement the propensity score method, we first estimate the distribution of delivery time as a function of order characteristics X_i.[31] In our main analysis, we start with the assumption that the delivery time follows a log normal distribution whose mean and variance are parameterized by the order characteristics X_i

$$log(D_i) \sim N(\theta(X_i), \sigma^2(X_i))$$

Because this is a prediction problem, we used standard machine learning toolkits to estimate $\widehat{\theta}(X_i)$ and $\widehat{\sigma}^2(X_i)$. Based on the estimated $(\widehat{\theta}(X_i), \widehat{\sigma}^2(X_i))$, we can then calculate the generalized propensity score function $\widehat{f}(D|X_i)$ based on the probability density that an order will arrive around minute D:

$$\widehat{f}(D|X_i) \equiv f_{LogNormal}(D|\widehat{\theta}(X_i), \widehat{\sigma}^2(X_i))$$

[31] This is different from the propensity score method in the binary case which only requires a measure of the treatment probability.

Similar to the standard propensity score, this generalized propensity score function allows us to reweight each delivery order observation based on the propensity that the order will arrive at a given minute. The weight is proportional to the inverse of this probability density:

$$\widehat{w}_i(D) = \frac{E[\widehat{f}(D|X)]}{\widehat{f}(D|X_i)} = \frac{\frac{1}{N}\sum_{j=1}^{N}\widehat{f}(D|X_j)}{\widehat{f}(D|X_i)}$$

This weight can then be used to estimate the average potential retention given delivery time D.

Another identification requirement is that sufficient variation in the treatment must exists such that the overlapping assumption is satisfied. Intuitively, if certain types of deliveries will never be late, then we cannot measure how customers may respond when such deliveries are late. This implies that we cannot study cases when delivery time deviates too far from promises because there are fewer observations that satisfy this criteria. We focus our analysis window from 60 minutes to 150 minutes to ensure that we have enough observations for orders that arrive at any given minute. We also rule out orders with limited variation in delivery time by removing orders that have less than 1% probability of taking less than 60 minutes or longer than 150 minutes.

B Reference points for experienced customers who used to receive earlier-than-promised deliveries

Another way to illustrate that reference point does not depend on past experience is to focus on a subset of experienced customers who used to receive delivery much earlier than promises, and test whether the reference point is still around promises. To examine this hypothesis, we focus on a subset of customers who have placed at least 5 orders and have experienced an average delivery time faster than 90 minutes in the past when they place their order. These customers should expect that the delivery will arrive much earlier than the promised delivery time of 2 hours. If these customers use such expectations as reference points, they will become more responsive to delivery time when the delivery takes longer than 90 minutes. Their responsiveness should not change around 2 hours. We estimate whether these customers change their responsiveness around 90 minutes by replicating the previous estimation methods on this subset of customers. Figure 18 visualizes the causal response for these subset of customers.

Table 5 formally estimates the reference point following Muggeo (2003). The estimated reference point is close to the promised delivery time but significantly different from 90 minutes, supporting the hypothesis that customers use promises as reference points.

Figure 18: Impact of the most recent delivery time on retention for customers who used to receive delivery more than 30 minutes earlier than promised

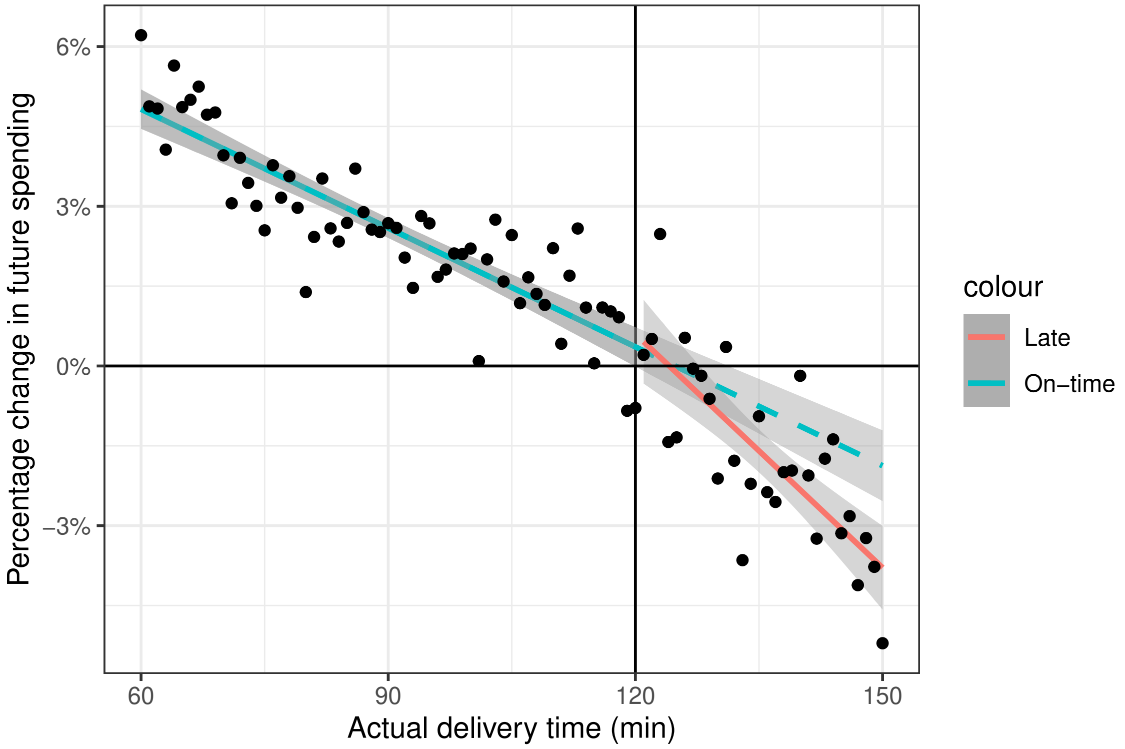


Table 5: Estimated reference point for customers who used to receive delivery more than 30 minutes earlier than promised

Outcome	Promises	Estimated Reference point	Confidence Interval CI(95%).low	 CI(95%).up
Future spending	120	123	112.417	133.583

Note: Estimated using segmented regression by Muggeo (2003)

The result supports the status-quo-based reference point primarily studied by Kahneman and Tversky (1979). The reference points are still close to the promise even if the average delivery time is much faster than promises.

C Alternative explanation to learning: heterogeneity and dynamic selection

Section 7.2 has demonstrated that the diminishing responsiveness is consistent with learning. It is possible that the difference in responsiveness is driven by customer heterogeneity and dynamic selection rather than learning: in-experienced customers include a subset of customers who are time-sensitive. These time-sensitive customers stop reordering after experiencing a late delivery. Therefore, customers who have placed 0 orders are on average more time sensitive than customers who have placed 30 orders. To address this concern of heterogeneity, we leverage the customer rating data as an outcome. Although rating and retention are closely related, rating reflects the current utility experienced by customers when they receive the order, while retention reflects the future utility expected by customers when they are about to reorder. If experienced customers are not time-sensitive, this insenstivity should also be reflected in their rating: the impact of delivery time on rating would be much smaller for experienced customers. We test this alternative explanation by measuring the causal response curve using rating rather than retention as the outcome. Figure 19 shows how the time sensitivity, measured by the impact of delivery time on rating, is related to customer experience. Both experienced and in-experienced customers are responsive to delivery time, suggesting that heterogeneity in time sensitivity cannot explain the diminishing pattern exhibited in Figure 9.

D Alternative model specifications

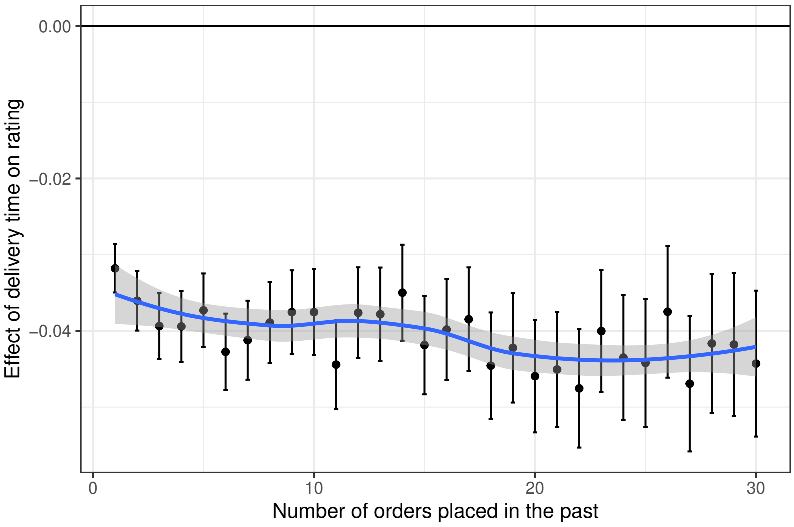
D.1 Bayesian learning with asymmetric preferences

This section discusses why the data pattern can not be rationalized by standard learning model with asymmetric preferences. Given a learning model, the effect of past delivery time on future retention can be decomposed into two components:

$$\frac{\partial \bar{U}_{t+1}}{\partial D_t} = \underbrace{\frac{\partial \bar{U}_{t+1}}{\partial \theta_{t+1}}}_{\substack{\text{Effect of belief} \\ \text{on expected utility}}} \times \underbrace{\frac{\partial \theta_{t+1}}{\partial D_t}}_{\substack{\text{Effect of delivery} \\ \text{time on belief}}}$$

where θ_{t+1} summarizes the posterior belief. Recall that Figure 4 and 5 have illustrated that customers responses are kinked around promises. Because the standard learning model with

Figure 19: Customer experience vs effect of delivery time on current rating



normal priors implies that the impact of delivery time on expected delivery time θ is smooth (linear):

$$\theta_{t+1} = (1-\lambda_t)\theta_t + \lambda_t(Promise_t - D_t),$$

the only way to rationalize the kinked response curve is to find a utility function such that the impact of delivery time on expected utility is kinked. However, this is impossible with standard parsimonious utility specification due to uncertainty. For example, if customers have quadratic preference for delivery time:

$$U_t = \alpha + \beta_1 D_t + \beta_2 D_t^2$$

The impact of belief on future expected utility is also going to be quadratic and smooth:

$$\begin{aligned}\bar{U}_{t+1} &= E[U_{t+1}(D_{t+1})|\theta_{t+1}, \sigma_{t+1}] \\ &= \alpha - \beta(Promise_{t+1} - \theta_{t+1}) - r[(Promise_{t+1} - \theta_{t+1})^2 + \sigma_{t+1}^2]\end{aligned} \tag{15}$$

As demonstrated by 5, this quadratic specification is less plausible compared to a piece-wise linear specification. Even if customers have a discontinuous plan-dependent utility:

$$u(D_t) = \begin{cases} \alpha - \beta D_t & D_t \leq P_t^{plan} \\ \alpha - \beta D_t - \gamma & D_t \geq P_t^{plan} \end{cases}$$

The expected utility is still smooth due to uncertainty:

$$\begin{aligned}\bar{U}_t = E[u(D_t)] &= \alpha - \beta(Promise_t - \theta_t) - \gamma E[D_t > P_t^{plan}] \\ &= \alpha - \beta(Promise_t - \theta_t) - \gamma\Phi(\frac{P_t^{plan} - (Promise_t - \theta_t)}{\sigma_t})\end{aligned}$$

Figure 20 illustrates the shape of the response for different utility specifications.

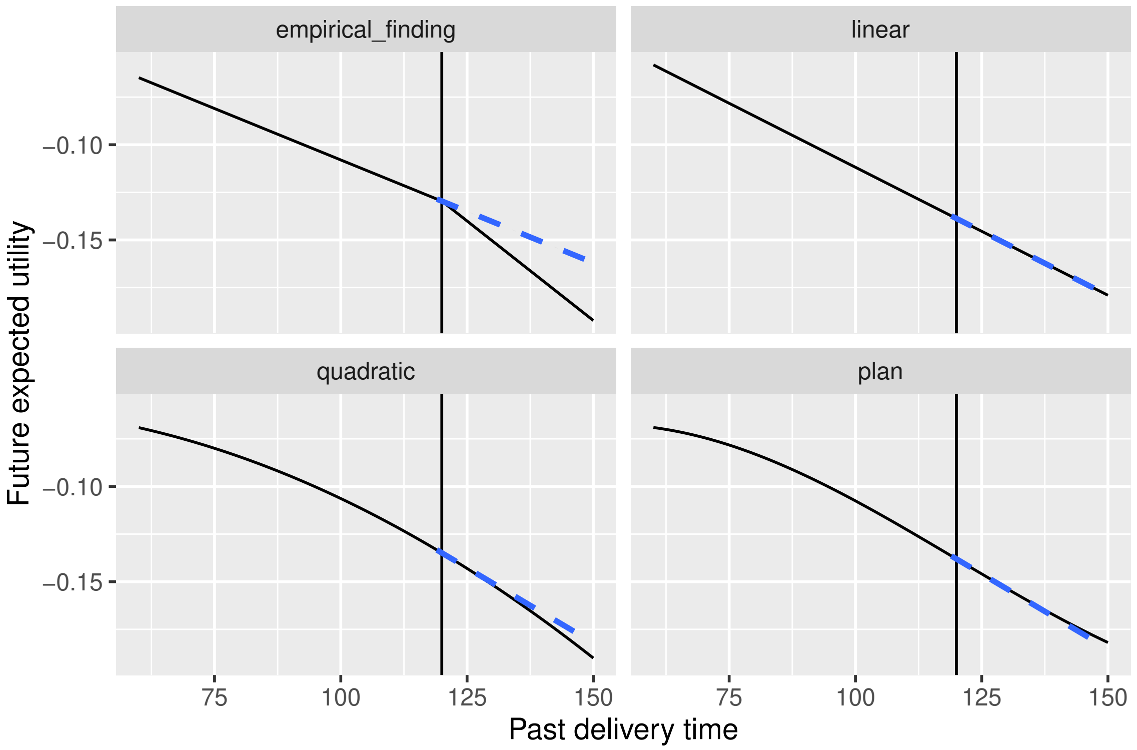


Figure 20: Shape of response curves under different utility specifications with Bayesian learning

D.2 A model of biased belief

Another way to rationalize the kinked response curve is to assume customers asymmetrically update their belief depending on whether the delivery is late or not:

$$\theta_{i,t+1} = \begin{cases} (1-\lambda_{it})\theta_{it} + \lambda_{it}(Promise_{it} - D_{it}) & D_{it} \leq Promise_{it} \\ (1-\lambda_{it}))\theta_{it} + \lambda_{it}(Promise_{it} - D_{it})(1+\eta) & D_{it} > Promise_{it} \end{cases}$$

where η captures the loss aversion when delivery arrives later than promises. Intuitively, if customers use promises as reference points and perceive late delivery as a loss, the delivery times will appear longer than they actually are, implying $\eta > 0$. Although this model can rationalize the change of slope, this model predicts that experienced customers will not be as responsive to late arrival as in-experienced customers, because $\lambda_{it}\eta$ is small for experienced customers. This is inconsistent with Figure 10 that demonstrates that late arrival still has a significant effect on experienced customers.

E Additional Figures and Tables

Table 6: Order characteristics: On-time vs late

Characteristics (X)	On-Time	Late
	$E[X_i \vert D_i <= 120]$	$E[X_i \vert D_i > 120]$
# past orders placed by customers	4.69	4.53
# items in the basket	18.09	19.49
Distance from grocery store (miles)	2.83	3.18
Is weekend	0.30	0.35

Note: To mask information that is business sensitive, all absolute characteristics are multiplied by a constant from $[0.5, 1.5]$.

Table 7: Estimated reference point

Outcome	Promises	Estimated	Confidence Interval	
		Reference point	CI(95%).low	CI(95%).up
Future spending	120	114.557	107.677	121.438

Note: Estimated using segmented regression by Muggeo (2003)

References

Ascarza, E., Netzer, O., and Hardie, B. G. S. (2018). Some Customers Would Rather Leave Without Saying Goodbye. *Marketing Science*, 37(1):54–77. Publisher: INFORMS.

Backus, M., Blake, T., Masterov, D., and Tadelis, S. (2021). Expectation, Disappointment, and Exit: Evidence on Reference Point Formation from an Online Marketplace. *Journal of the European Economic Association*, (jvab033).

Barberis, N. C. (2013). Thirty Years of Prospect Theory in Economics: A Review and Assessment. *Journal of Economic Perspectives*, 27(1):173–196.

Bell, D. R. and Lattin, J. M. (2000). Looking for Loss Aversion in Scanner Panel Data: The Confounding Effect of Price Response Heterogeneity. *Marketing Science*, 19(2):185–200. Publisher: INFORMS.

Bolton, R. N. (1998). A Dynamic Model of the Duration of the Customer's Relationship with a Continuous Service Provider: The Role of Satisfaction. *Marketing Science*, 17(1):45–65. Publisher: INFORMS.

Braun, M. and Schweidel, D. A. (2011). Modeling Customer Lifetimes with Multiple Causes of Churn. *Marketing Science*, 30(5):881–902. Publisher: INFORMS.

Cattaneo, M. D., Crump, R. K., Farrell, M. H., and Feng, Y. (2021). On Binscatter. Number: arXiv:1902.09608 arXiv:1902.09608 [econ, stat].

Gijsenberg, M. J., Van Heerde, H. J., and Verhoef, P. C. (2015). Losses Loom Longer than Gains: Modeling the Impact of Service Crises on Perceived Service Quality over Time. *Journal of Marketing Research*, 52(5):642–656. Publisher: SAGE Publications Inc.

Gill, D. and Prowse, V. (2012). A Structural Analysis of Disappointment Aversion in a Real Effort Competition. *The American Economic Review*, 102(1):469–503. Publisher: American Economic Association.

Gneezy, A. and Epley, N. (2014). Worth Keeping but Not Exceeding: Asymmetric Consequences of Breaking Versus Exceeding Promises. *Social Psychological and Personality Science*, 5(7):796–804. Publisher: SAGE Publications Inc.

Heffetz, O. and List, J. A. (2014). IS THE ENDOWMENT EFFECT AN EXPECTATIONS EFFECT?: Is the Endowment Effect an Expectations Effect? *Journal of the European Economic Association*, 12(5):1396–1422.

Hirano, K. and Imbens, G. W. (2005). The Propensity Score with Continuous Treatments. In Gelman, A. and Meng, X.-L., editors, *Wiley Series in Probability and Statistics*, pages 73–84. John Wiley & Sons, Ltd, Chichester, UK.

Ho, T. H. and Zheng, Y.-S. (2004). Setting Customer Expectation in Service Delivery: An Integrated Marketing-Operations Perspective. *Management Science*, 50(4):479–488. Publisher: INFORMS.

Huang, G. and Liu, H. (2021). Estimating expectations-based reference-price effects in the used-car retail market. *Quantitative Marketing and Economics*.

Imai, K. and van Dyk, D. A. (2004). Causal Inference With General Treatment Regimes: Generalizing the Propensity Score. *Journal of the American Statistical Association*, 99(467):854–866.

Joshi, Y. V. and Musalem, A. (2012). Underpromising and Overdelivering: Strategic Implications of Word of Mouth. SSRN Scholarly Paper 1945970, Social Science Research Network, Rochester, NY.

Kahneman, D. and Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2):263–291. Publisher: [Wiley, Econometric Society].

Kim, Y. (2021). Customer Retention under Imperfect Information. SSRN Scholarly Paper ID 3709043, Social Science Research Network, Rochester, NY.

Klemperer, P. (1995). Competition when Consumers have Switching Costs: An Overview with Applications to Industrial Organization, Macroeconomics, and International Trade. *The Review of Economic Studies*, 62(4):515–539. Publisher: [Oxford University Press, Review of Economic Studies, Ltd.].

Kopalle, P. K. and Lehmann, D. R. (2001). Strategic Management of Expectations: The Role of Disconfirmation Sensitivity and Perfectionism. *Journal of Marketing Research*, 38(3):386–394. Publisher: SAGE Publications Inc.

Kopalle, P. K. and Lehmann, D. R. (2006). Setting Quality Expectations When Entering a Market: What Should the Promise Be? *Marketing Science*, 25(1):8–24. Publisher: INFORMS.

Kumar, P., Kalwani, M. U., and Dada, M. (1997). The Impact of Waiting Time Guarantees on Customers' Waiting Experiences. *Marketing Science*, 16(4):295–314. Publisher: INFORMS.

Kőszegi, B. and Rabin, M. (2006). A Model of Reference-Dependent Preferences*. *The Quarterly Journal of Economics*, 121(4):1133–1165.

Martin, S. and Shelegia, S. (2021). Underpromise and overdeliver? - Online product reviews and firm pricing. *International Journal of Industrial Organization*, 79:102775.

Marzilli Ericson, K. M. and Fuster, A. (2011). Expectations as Endowments: Evidence on Reference-Dependent Preferences from Exchange and Valuation Experiments *. *The Quarterly Journal of Economics*, 126(4):1879–1907.

Muggeo, V. M. R. (2003). Estimating regression models with unknown break-points. *Statistics in Medicine*, 22(19):3055–3071. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.1545.

Parasuraman, A., Zeithaml, V. A., and Berry, L. L. (1985). A Conceptual Model of Service Quality and Its Implications for Future Research. *Journal of Marketing*, 49(4):41–50. Publisher: SAGE Publications Inc.

Parasuraman, A. P., Zeithaml, V., and Berry, L. (1988). SERVQUAL: A multiple- Item Scale for measuring consumer perceptions of service quality. *Journal of retailing*.

Peters, T. (1988). *Thriving on Chaos: Handbook for a Management Revolution*. HarperCollins. Google-Books-ID: KIKNNWYcCt0C.

Sewell, C. and Brown, P. B. (2009). *Customers for Life: How to Turn That One-Time Buyer Into a Lifetime Customer*. Crown. Google-Books-ID: 5SiVOIfcEDgC.

Spence, M. (1977). Consumer Misperceptions, Product Failure and Producer Liability. *The Review of Economic Studies*, 44(3):561–572. Publisher: [Oxford University Press, Review of Economic Studies, Ltd.].

Sriram, S., Chintagunta, P. K., and Manchanda, P. (2015). Service Quality Variability and Termination Behavior. *Management Science*, 61(11):2739–2759.