When Talk is "Free": The Effect of Tariff Structure on Usage under Two- and Three-Part Tariffs

Eva Ascarza, Anja Lambrecht, and Naufel Vilcassim

Web Appendix

In this appendix, we present a detailed description of the analyses performed to obtain certain results discussed in the main manuscript.

DESCRIPTIVE ANALYSES

Analysis of tariff choice

We analyze whether customers' bills would have been lower on another than their chosen tariff at the time that three-part tariffs were introduced. Based on the three available usage periods prior to the three-part tariff introduction, we compute the individual-level average usage and standard deviation. For simplicity, we exclude customers who have switched more than once as well as the 1.1% of customers who switched within these three months.

To account for deviations from average usage due to random usage shocks, we then compute the bill for the usage level of [average usage +/− 1 standard deviation] under the current tariff, and the bill for the average usage under each of the remaining tariffs. We conclude that a customer would have had a lower bill on a different tariff if the bill for their average usage on a tariff other than the chosen tariff was below the lower bound of the bill-interval that accounts for variation in usage on the chosen tariff. Note that this analysis focuses on potential savings and does not account for the fact that customers may, on the same bill, be able to use more on a different tariff. The next section will discuss this aspect in detail.

Table A1 illustrates that based on their average usage and standard deviation of usage before the introduction of three-part tariffs, the large majority of customers chose the tariff that

minimizes their bill. In total, 26.2% of customers would pay less on a different tariff. For customers that would pay less on a different tariff average savings were between MU 4.1 and MU 7.7. As a result, it would take customers more than one period on average to amortize the switching fee of MU 10.

We then exclude three-part tariffs from this analysis and limit the analysis to whether customers would have paid less on a different two-part tariff. We find that only 10.9% of customers would have paid less on a different two-part tariff. This further confirms that two-part tariff customers largely chose the bill-minimizing tariff.

Table A1: Potential savings when three-part tariffs were introduced

	Tariff with lowest bill (in %)								
Chosen tariff	T_2_1	T_2_2	T_2_3	T_2_4	T_3_1	T_3_2	T_3_3	N	Avg. savings (in MU)*
Tariff_2_1	78.8	5.5	0.0	0.0	13.4	1.7	0.6	850	7.7
Tariff_2_2	0.3	87.0	0.0	0.0	9.1	1.8	1.8	814	6.3
Tariff_2_3	2.5	3.7	72.5	0.0	13.9	4.8	2.7	2,815	5.7
Tariff_2_4	1.4	9.4	0.0	65.0	12.4	6.2	5.6	1,253	4.1

Excluding customers who switched within the first 3 months of our data and customers that in our data switch more than once

* Average savings on tariff with lowest bill, computed only for customers that would have had a lower bill on a different tariff

Detailed analysis of switching from two- to three-part tariffs

The previous section focused on whether customers would have *paid less* on a different tariff. We now focus on three-part tariffs and analyze in more detail whether customers would benefit from switching to a three-part tariff, accounting for both whether customers would have *paid less* on a different tariff and whether they would have been able to *use more* for the same bill.

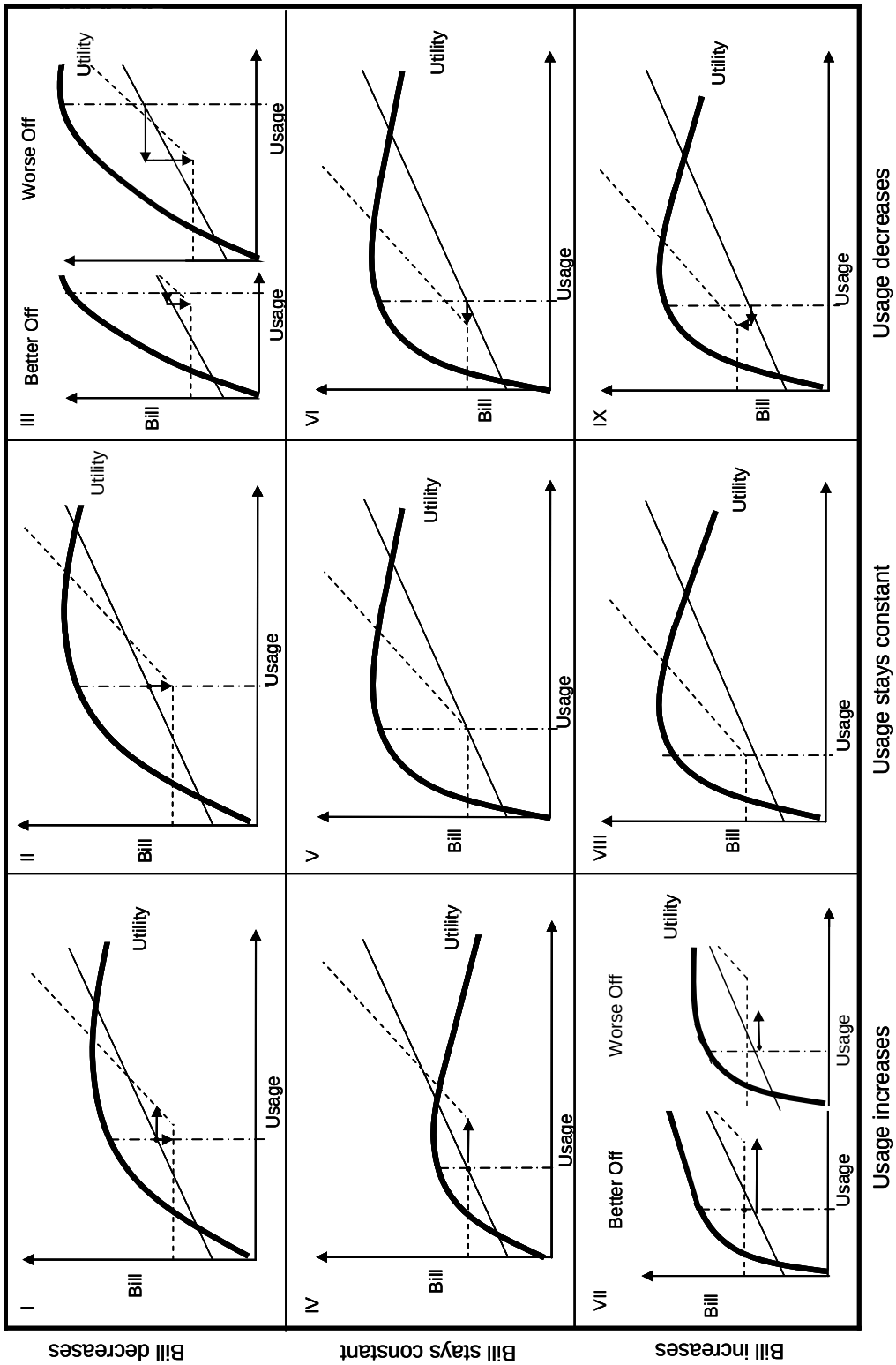
Figure A1 illustrates in which situations a customer should or should not switch to a three-part tariff. We abstract from switching costs and assume that a customer knows her optimal usage under a two- and a three-part tariff. We assume a utility function which is quadratic in usage (bold curve; the Model section of the main paper justifies the choice of utility function). The bill on a two-

part tariff (straight line) increases in the customer's usage. The bill on a three-part tariff (dashed line) remains flat as long as usage remains within the allowance and then increases linearly in usage. The maximum distance between the utility function and the bill indicates a customer's surplus on that tariff. A rational customer should switch to a three-part tariff if that entails a greater surplus than on a two-part tariff.

The vertical (horizontal) arrows indicate how such a switch would affect a customer's bill (usage). A customer should switch to a three-part tariff if for the same optimal usage, she pays less on a three- than on a two-part tariff (II), for the same bill, her optimal usage is greater on a three- than on a two-part tariff (IV), or if she can increase her optimal usage and still pay less on a three-part tariff (I). A customer should not switch if for the same bill, her optimal usage on a three-part tariff would decrease (VI), for the same optimal usage her bill would increase (VIII) or if her bill would increase while decreasing optimal usage (IX). She is indifferent if the same optimal usage entails the same bill (V). If under a three-part tariff, both optimal usage and the bill would decrease (III) or increase (VII), switching may or may not be beneficial, depending on the curvature of the utility function.

To determine which customers in our sample should or should not switch to a three-part tariff, we compare actual usage and expenditures on a two-part tariff to (a) how much a customer could use under a three-part tariff for the same bill and (b) how much she would pay under a three-part tariff for the same usage (Figure A1). To account for deviations from average usage due to random usage shocks, the interval of [average usage +/− 1 standard deviation] and the interval of the bill of [average usage +/− 1 standard deviation] serve as a reference point. For example, we classify a customer as being indifferent between switching to a three-part tariff and staying on a two-part tariff (Case V) if the same optimal usage entails a bill in the same interval on a two- and a three-part tariff.

Figure A1: Predicted switching from two- to three-part tariffs



The maximum distance between the utility function and a tariff's bill indicates maximum surplus.
The vertical dotted line represents the optimal level of usage on a two-part tariff.
Note: Marginal utility and the price of the outside good are set to 1, so utility represents willingness to pay.

Table A2 summarizes the results of this analysis. The first four columns correspond to the results when the switching fee is not taken into consideration. They indicate that customers who, according to our analysis, should switch to a three-part tariff were far more likely to switch to a three-part tariff than customers who according to our analysis should not switch to a three-part tariff. The next set of results accounts for the fee the customer has to pay for switching. Here we consider a switch to be beneficial if savings in the first month would compensate for the switching fee. Since the fee increases the bill, the share of customers classified as "unknown", i.e., those for whom both optimal usage and the bill would increase on a three-part tariff, is larger than when abstracting from the switching fee.

Table A2: Predicted and actual switching behavior

	Not considering switching fee				**Considering switching fee**			
Category	No. of customers	% of sample	% of customers in that group who switched	% of total switchers belonging to category	No. of customers	% of sample	% of customers in that group who switched	% of total switchers belonging to category
Should switch	3,710	63.7%	8.95%	71.7%	980	16.8%	13.2%	27.9%
Should not switch	85	1.5%	4.71%	0.9%	704	12.1%	6.4%	9.7%
Indifferent[(a)]	765	13.1%	7.19%	11.9%	211	3.6%	10.4%	4.8%
Unknown[(b)]	1271	21.8%	5.66%	15.6%	3936	67.5%	6.8%	57.7%

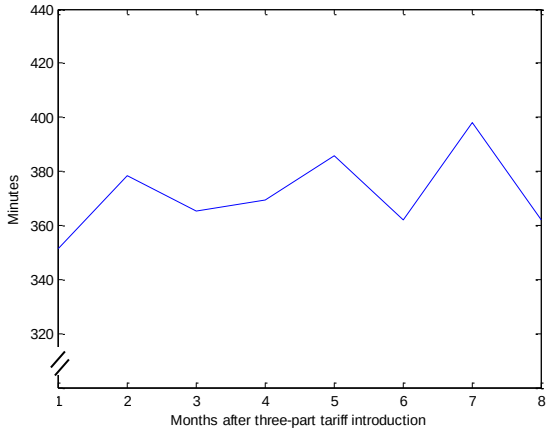
(a) A customer is indifferent if the same optimal usage entails the same bill on a two- and a three-part tariff.

(b) If under a three-part tariff, both optimal usage and the bill would decrease or increase, switching may or may not be beneficial depending on the curvature of the utility function.

Persistence of three-part tariff usage over time

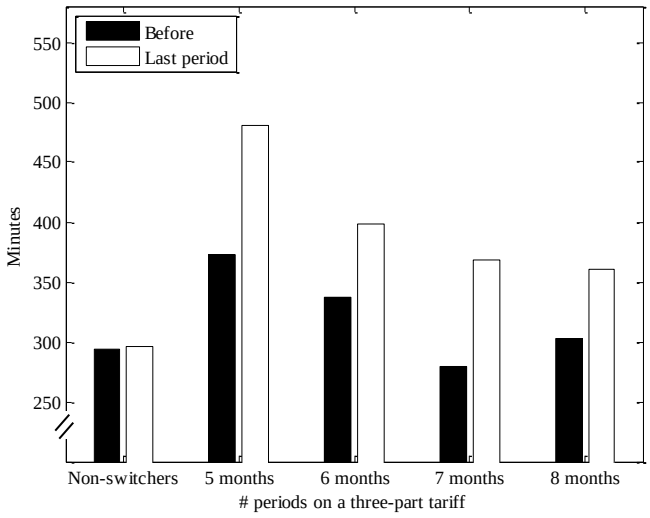
We next check whether the increase in three-part tariff usage persists over time. We focus on customers for whom we observe at least six months of three-part tariff usage and plot the aggregate three-part tariff usage over time. Figure A2 illustrates that, apart from the holiday seasons in months 5 and 7 after the introduction of the three-part tariffs, there are no clear trends of increasing or decreasing three-part tariff usage.

Figure A2: Monthly average usage after switching to a three-part tariff



Second, we compare average usage before and after the three-part tariff introduction, as we do in the Descriptive Analysis section of the main manuscript, but now analyze differences by cohorts (i.e., groups of customers who switched to a three-part tariff in the same month). Figure A3 shows, for each cohort, the average usage before the three-part tariffs were introduced and the average usage in the last period of our data and compares it to customers who did not switch to a three-part tariff. We observe a consistent increment in usage among three-part tariff switchers, regardless of how long customers have been on a three-part tariff.

Figure A3: Average usage before and after the introduction of three-part tariffs, by cohorts

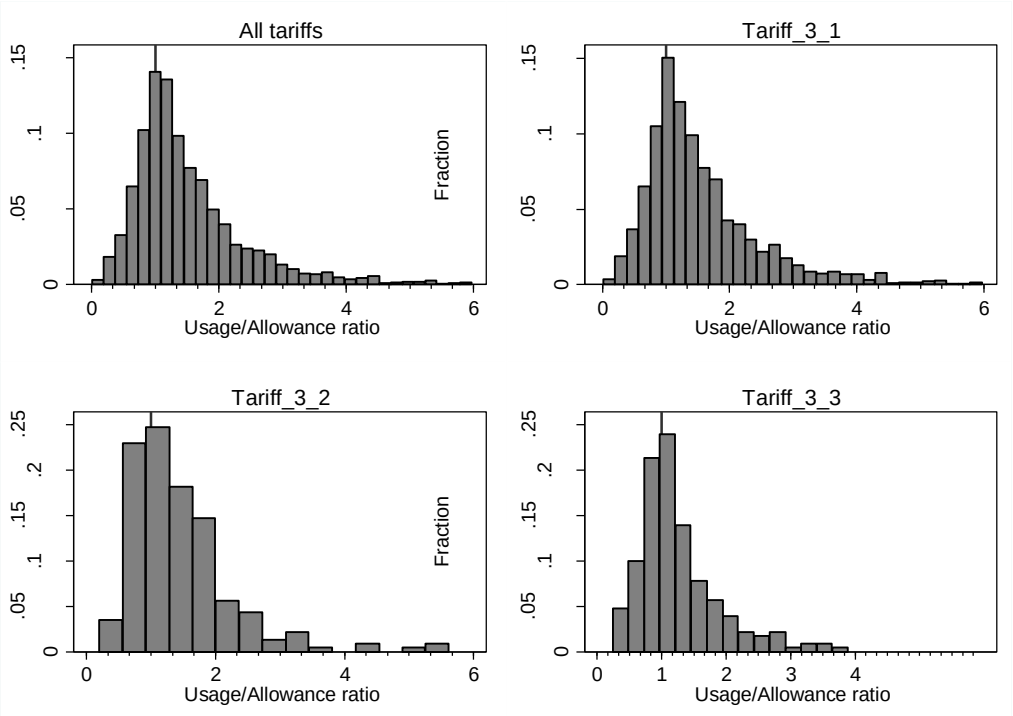


Further details on three-part tariff usage behavior

We summarize information on customers' usage behavior on three-part tariffs. First, we analyze the distribution of usage as percentage of the allowance (Figure A4). Across all three-part tariffs, we observe a mass point of usage observations when usage is approximately equal to the allowance. This mass point results from the type of budget constraint imposed by a three-part tariff that implies bunching of usage observations at 100% of the allowance (see equation (4) in the main manuscript). It is reassuring that we indeed find such a mass point in our data since it provides additional evidence that customers are aware of their usage behavior.

Figure A4 also illustrates that many customers use more than their usage allowance. This is in line with the behavioral motivation that leads to greater three-part tariff usage as discussed in the main paper. It outlines that the positive effect from a three-part tariff should persist when consumers have exceeded their allowance.

Figure A4: Usage as a percent of allowance



Second, we analyze whether three-part tariff customers had chosen the ex-post bill-minimizing tariff based on their first three months of three-part tariff usage. As in the first section of this web appendix, we rely on the bill for the usage level of [average usage +/- 1 standard deviation] under the current tariff, and the bill for the average usage under each available tariff. Table A3 illustrates that overall 86.8% of customers chose the three-part tariff that minimizes their bill based on their *ex post* usage. Since the differences between access prices and allowances between the three-part tariffs are large, even customers that use more than their allowance are largely in the bill-minimizing tariff.

Table A3: Optimality of chosen three-part tariff (based on first three periods on a three-part tariff)

	Tariff with lowest bill (in %)				
Chosen tariff	Two-part tariff	T_3_1	T_3_2	T_3_3	N
Tariff_3_1	5.7	86.8	5.1	2.4	296
Tariff_3_2	0.0	6.1	81.8	12.1	33
Tariff_3_3	0.0	0.0	8.8	91.2	34

Includes all customers with at least three periods on a three-part tariff, excludes customers who switched again in their first three periods on a three-part tariff

DEMAND ESTIMATION

Linear demand estimation of three-part tariff usage

We compare actual usage on two- and three-part tariffs to predicted usage for the last month in our data. We estimate a linear demand function for two-part tariff usage, $q_{ijt} = d_{it} - bp_j$, where q_{ijt} denotes the number of minutes that individual *i* consumes on tariff *j* at time *t*, d_{it} denotes the satiation level, or demand intercept, b refers to the price coefficient and p_j is the usage price of tariff *j*. Since we have little within-customer variation of the usage price, the price coefficient is assumed to be homogenous across customers. We incorporate an individual-level preference, η_i, and a multiplicative shock, ϕ_{it}, into the demand intercept, $d_{it} = \phi_{it} e^{\eta_i}$. We assume that η_i follows a

normal distribution with mean and variance $\left(\mu_\eta, \sigma_\eta^2\right)$ and that ϕ_{it} is distributed lognormal with parameters $\left(-0.5\sigma_\phi^2, \sigma_\phi^2\right)$, such that $E(\phi_{it}) = 1$. We use MCMC methods to estimate the model. We choose diffuse hyperpriors for b, μ_η, σ_η, and σ_ϕ. We burn-in 90,000 iterations and use the next 10,000 to sample from the posterior distributions of the parameters of interest and to predict consumption in the last period of data. The parameters estimates are shown in Table A4.

Table A4: Estimation Results (Homogeneous price coefficient)

	Mean	95% Interval	
b	421.158	371.841	478.932
μ_η	5.519	5.498	5.545
σ_η	0.679	0.662	0.696
σ_ϕ	0.308	0.300	0.316

For customers who remained on a two-part tariff, we predict consumption in the last period of the data as:

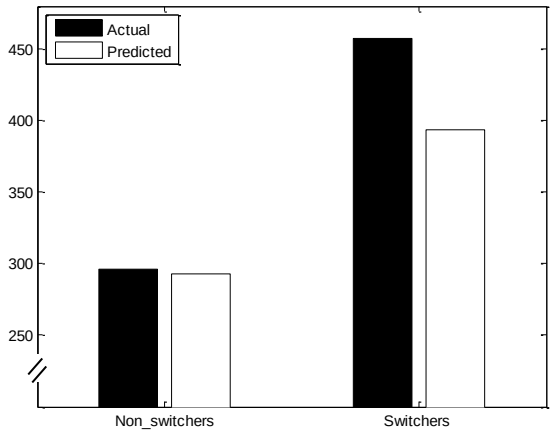
$$q_{ijt}^* = \begin{cases} 0 & \text{if } d_{ijt} \le bp_j \\ d_{ijt} - bp_j & \text{if } d_{ijt} > bp_j . \end{cases}$$

For customers who have switched to a three-part tariff, we predict consumption in the last period of the data as

$$q_{ijt}^* = \begin{cases} d_{ijt} & \text{if } d_{ijt} < \tilde{q}_j \\ \text{Max}(\tilde{q}_j, d_{ijt} - bp_j) & \text{if } d_{ijt} > \tilde{q}_j \end{cases}$$

Figure A5 illustrates that the model accurately predicts usage for customers who remain on a two-part tariff while notably underpredicting consumption for customers who switch to a three-part tariff. In other words, the model does not capture the increment in usage observed for three-part tariff customers.

Figure A5: Usage predictions using linear model (all customers)



Persistence of over-usage over time

We next check whether the unpredicted increase in three-part tariff usage persists over time. We use the estimates obtained in the analysis presented in the previous section but now analyze three-part tariff customers in cohorts of customers who switched to a three-part tariff in the same month. For each cohort, we predict usage in the last month of the data and compare it with actual usage in that month. The model under-predicts three-part tariff usage regardless of how long customers have been on a three-part tariff. Specifically, we under-predict usage by 22.1% for the five-month cohort, by 12.7% for the six-month cohort, by 19.8% for the seven-month cohort, and by 12.1% for the eight-month cohort.

Robustness to non-linear demand specifications

If customers' usage followed a convex demand function, our linear demand model in the previous section would predict demand accurately in the area of usage prices similar to those of the two-part tariffs, i.e., 0.042–0.079 MU, but would possibly underpredict usage at a zero price. As a consequence, the over-usage we find in the descriptive analysis presented in the main manuscript could simply be due to the specification of the demand function. We rule out this possibility by estimating two additional demand specifications.

First, we use a polynomial specification (as a Taylor approximation to the true demand function) to estimate demand. We build on the demand function presented in the previous section, $q_{ijt} = d_{it} - bp_j$, and include a quadratic term, $b_2 p_j^2$, and a cubic term $b_3 p_j^3$. We estimate demand as $q_{ijt} = d_{it} - b_1 p_j - b_2 p_j^2 - b_3 p_j^3$. If the quadratic and cubic terms do not significantly differ from zero, that would support the choice of a linear demand function.

We replicate the analysis presented in the main manuscript. The results show that the quadratic and cubic terms of the demand function are not significantly different from zero (Table A5).

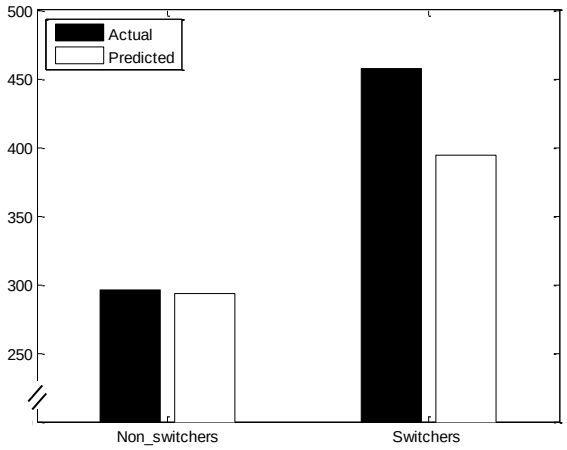
We next use the parameter estimates to predict usage in the last period.

Figure A6 displays the results. Similarly to our main specification, predicted usage of customers who switched to a three-part tariff is only 86.4% of their actual usage while the model predicts 98.9% of actual usage for customers who remain on a two-part tariff. This provides evidence that the increase in usage is not due to the specific form of the demand function.

Table A5: Estimation results (quadratic and cubic terms)

	Mean	95% Posterior Interval	
μ_η	5.519	5.504	5.534
σ_η	0.674	0.659	0.689
b_1	442.398	396.798	486.488
b_2	4.742	-49.264	60.743
b_3	-1.984	-63.673	62.446
σ_ϕ	0.306	0.299	0.315

Figure A6: Usage predictions using quadratic and cubic terms



Second, we estimate an additional model specification that allows for convex demand: $q_{ijt} = \frac{e^{\eta_i + \nu_{it}}}{p_j - \gamma} - \beta$. This demand specification is obtained by maximizing the utility function $U_{ijt}(q_{ijt}, q_{i0t}) = \alpha \log(q_{ijt} + \beta) + \gamma q_{ijt} + q_{i0t}$, with $\alpha, \beta > 0$ and $\gamma < 0$. The term q_{i0t} denotes the outside good, when its price is being normalized to 1.

To empirically disentangle η_i and , γ the data needs to have individual-level variation of the usage price. However, in our data there is little tariff switching before the three-part tariffs were introduced. An alternative is to fix the value of γ at a reasonable level and estimate the remaining parameters based on the first two periods and predict usage for the last period. We proceed in three steps:

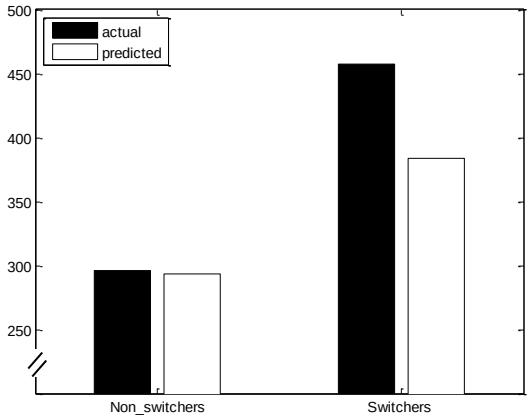
1. To avoid having to arbitrarily set γ, we estimate the demand model using all observations from the first six periods of data. We obtain an estimate of γ (-0.049).[1]

[1] We conduct two sets of robustness checks to our estimate of γ. First, we estimate γ based on a different number of periods (4 and 6 periods). Second, we estimate γ based on a random subsample of 50% of the customers in our dataset. We find that our estimate of γ is robust to these alternative specifications.

2. We then set γ and estimate the remaining parameters, including η_i, using two-part tariff usage observations prior to the three-part tariff introduction.
3. We then use the set of estimated parameters to predict usage in the last period of our data.

Figure A7 illustrates predicted versus actual usage. Consistent with the results obtained in the previous section, we under-predict three-part tariff usage by 19.2% while predicting two-part tariff usage very accurately (under-prediction of only 0.9%). This provides further evidence that the assumption of linear demand does not lead us to artificially under-predict three-part tariff usage.

Figure A7: Usage prediction convex demand function



Robustness to non-homogeneous price sensitivity

It is possible that customers who switch to a three-part tariff differ in their usage price sensitivity from customers who remain on a two-part tariff. Given the limited within-customer price variation in our data, we cannot estimate a model with an individual-level price coefficient, b_i. Nevertheless, we conduct an *ad hoc* analysis in which we allow for a different set of parameters for switchers to a three-part tariff compared to all other customers. We then test whether this specification still under-predicts three-part tariff usage.

As in the Descriptive Analysis section in the main manuscript, we estimate a demand model using the two-part tariff periods prior to the three-part tariff introduction and then predict usage in

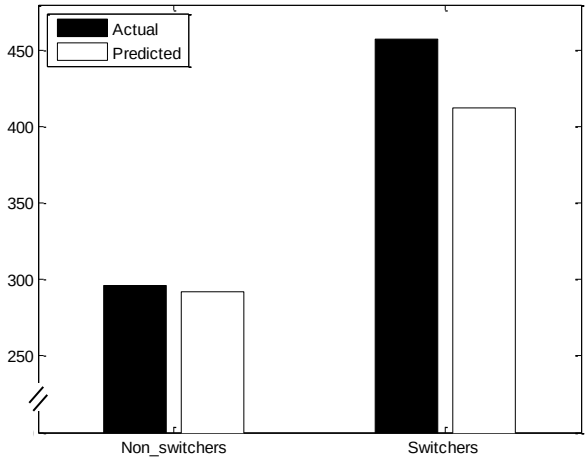
the last period of our data. We now estimate two sets of coefficients, one for customers who remained on a two-part tariff and one for customers who switched to a three-part tariff. The same diffuse priors were chosen for both sets of parameters. Table A6 summarizes the posterior distributions and Figure A8 shows the model predictions.

The model with heterogeneity in price sensitivity under-predicts three-part tariff usage by 9.8% while two-part tariff usage is predicted very accurately. We conclude that while heterogeneity in usage price sensitivity may possibly contribute to greater three-part tariff usage, it does not explain the large increase in usage we observe in the data. To capture some degree of heterogeneity in usage price sensitivity, our full model specification (the Model section of the main manuscript) incorporates observed heterogeneity in the usage price sensitivity.

Table A6: Estimation results (heterogeneous price coefficient)

	Customers who do not switch to a three-part tariff			Customers who switch to a three-part tariff		
	Mean	95% Interval		Mean	95% Interval	
b	422.556	368.723	476.531	764.364	542.254	1027.587
μ_η	5.501	5.478	5.525	5.817	5.733	5.895
σ_η	0.668	0.651	0.686	0.700	0.649	0.739
σ_ϕ	0.307	0.299	0.314	0.294	0.268	0.323

Figure A8: Usage predictions for heterogeneity in price sensitivity (linear model)



Analysis of autocorrelation in the usage process leading to self-selection

As discussed in the main manuscript, autocorrelation in the usage process could be a possible explanation for the usage increase we observe. If usage followed an autoregressive process and customers switched to a three-part tariff after having received a positive usage shock, then we would expect that customers increase their consumption after switching to a three-part tariff. However, we find that this pattern of behavior is not consistent with our data.

We first investigate the level of autocorrelation among the usage shocks. Given that our demand is specified with multiplicative usage shocks in the demand coefficient, shocks do not enter in a linear way. Hence, we cannot run simple autocorrelation tests using usage observations. To isolate the usage shocks, one would need to take logs of the quantity $\left(q_{ijt} + bp_j\right)$, which is not feasible since b is one of the parameters to be estimated. To overcome this issue, we consider sub-samples of customers for which p_j does not vary, reducing the term pb_j to a constant, and then estimate the degree of autocorrelation in each sub-sample. We do so by successively limiting the sample to customers who are on the same tariff and do not switch to a different tariff. Then we run

a fixed effect linear regression for the whole history of each set of customers, using $\log(q_{ijt})$ as dependent variable and its lagged value as independent variable.[2] For each of the subset of customers, we find no evidence of strong autocorrelation among the usage shocks (ρ ranges from 0.16 to 0.35 across all tariffs).

We then perform further analyses to ensure that the weak serial correlation we find does not bias our model estimates. We first simulate tariff choice and usage behavior for a synthetic panel of customers where we use the estimated parameters from our main model as the data generating process. We incorporate weak autocorrelation (values of 0.2, 0.3 and 0.4) into the usage process through autocorrelated usage shocks. We estimate all parameters using our main model. We find that in all cases the simulated values lay within the posterior interval of the estimated parameters. This provides further confirmation that our results are not affected by a possible weak autocorrelation.

Second, we investigate whether past usage shocks affect switching behavior. We estimate a logistic regression with 'switching to a three-part tariff' as dependent variable.[3] As independent variables, we use past usage, dummy variables for the current two-part tariff, and the ratio of usage in the last period to usage in the period before last. The latter variable serves as a proxy for the usage shock received in the previous period. If past usage shocks affected switching to three-part tariffs, then the "shock" variable should be significant. We find that this is not the case.

Table A7 summarizes the results of three different specifications. In the first specification, we include the usage shock in the last period as a predictor for switching behavior, controlling for the chosen tariff. In the second specification, we also control for the average usage level previous to

---

[2] We use the method proposed by Blundell and Bond (1998) to correct for the Nickell bias induced by the fixed effect.

[3] We estimate tariff choice in the fourth month of data. As a robustness check we also estimate the same model using months 5, 6, etc. and in all cases, obtain qualitatively the same results.

the three-part tariff introduction, and in the third specification, we add a quadratic term for average usage.[4] In all specifications, the proxy for a past usage shock is not significant. We therefore conclude that autocorrelation does not explain the over-usage we observe in the data.

Table A7: Logistic regression results for switching to three-part tariffs

Variable	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value
Constant	-5.020	0.000	-5.496	0.000	-5.582	0.000
Previous usage (avg.)			0.001	0.008	0.001	0.130
Previous usage (avg.) ^2					0.000	0.645
Past usage shock	0.034	0.206	0.034	0.226	0.035	0.233
Dummy for previous tariff 2_1	0.635	0.160	0.966	0.045	1.000	0.039
Dummy for previous tariff 2_2	-1.037	0.186	-0.763	0.338	-0.746	0.349
Dummy for previous tariff 2_3	0.755	0.041	0.978	0.012	0.988	0.011

MODEL

Asymptotic properties of the learning model

We show that for any value of the initial parameters (α_0, β_0), the expected value of the belief $\tilde{\beta}_i$ converges to the true value, β_i, and its variance goes to zero as the consumer gets more experience on a three-part tariff (i.e., the number of periods on a three part tariff goes to infinity). We compute the limit of the mean and the variance of the beliefs, as shown in equation (22), when n goes to infinity:

$$\lim_{n\to\infty} E(\tilde{\beta}_{i\tau_n}) = \lim_{n\to\infty} \frac{\alpha_0 + nr}{\beta_0 + \sum_{t=1}^{n} s_{i\tau_t}}$$

(A1)

$$= \lim_{n\to\infty} \frac{\frac{\alpha_0}{n} + r}{\frac{\beta_0}{n} + \frac{1}{n}\sum_{t=1}^{n} s_{i\tau_t}}.$$

[4] We perform the same analysis using (1) current usage, and (2) lagged usage. We obtain the same qualitative results.

We know from equation (19), that $s_{i\tau_t}$ is gamma-distributed with shape and scale parameters $\left(r, r/e^{\delta_i}\right)$. Thus, as $n \to \infty$, we know that $\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} s_{i\tau_t} = e^{\delta_i}$.

Therefore, substituting this result into (A1), we obtain that $\lim_{n\to\infty} E(\tilde{\beta}_{i\tau_n}) = \frac{r}{e^{\delta_i}}$.

$$\lim_{n\to\infty} Var(\tilde{\beta}_{i\tau_n}) = \lim_{n\to\infty} \frac{\alpha_0 + nr}{\left(\beta_0 + \sum_{t=1}^{n} s_{i\tau_t}\right)^2}$$

$$= \lim_{n\to\infty} \frac{\frac{\alpha_0}{n} + r}{n\left(\frac{\beta_0}{n} + \frac{1}{n}\sum_{t=1}^{n} s_{i\tau_t}\right)^2} \quad \text{(A2)}$$

$$= 0.$$

Posterior distributions for the full model

The model is estimated using a Bayesian framework. We obtain estimates of all model parameters by drawing from the marginal posterior distributions. Given the nonlinearities of our likelihood function and the complexity of the hierarchy in the parameters, most conditional distributions do not have conjugate posteriors. We use the Metropolis-Hasting (MH) algorithm to draw from these conditional posterior distributions. We use a data augmentation approach to include the unobserved individual-level parameters as well as the time-variant beliefs.

We denote Ω as all parameters in our model, including the population parameters $\Phi = \{b, \rho_1, \rho_2, \beta_0, a_1, a_2, \exp(r)\}$, the individual-level parameters $\varpi_i = \{\eta_i, \delta_i, \lambda_i\}$, the mixing parameters $\alpha = \{\mu_\eta, \sigma_\eta, \mu_\delta, \sigma_\delta, \mu_\lambda, \sigma_\lambda\}$, and the individual specific time-variant beliefs $\tilde{\beta}_{it}$. The full joint posterior distribution can be written as:

$$
\begin{aligned}
f(\Omega \mid Data) &\propto L(Data \mid \Omega) f(\Omega) \\
&= \left\{ \prod_{i=1}^{I} \left\{ \prod_{t=1}^{T_i} f(q_{it} \mid k_{it}, \Phi, \eta_i, \delta_i, Z_{it}, X_j) \right. \right. \\
&\qquad \times f(k_{it} \mid \tilde{\beta}_{it}, \Phi, \eta_i, \delta_i, \lambda_i, Z_{it}, X_j) \\
&\qquad \left. \times f(\tilde{\beta}_{it} \mid \Phi, \eta_i, \beta_0, Z_{it}) \right\} \\
&\quad \left. \times f(\eta_i \mid \mu_\eta, \sigma_\eta) f(\delta_i \mid \mu_\delta, \sigma_\delta) f(\lambda_i \mid \mu_\lambda, \sigma_\lambda) \right\} \\
&\quad \times f(\mu_\eta) f(\sigma_\eta) f(\mu_\delta) f(\sigma_\delta) f(\mu_\lambda) f(\sigma_\lambda) f(\Phi).
\end{aligned}
$$

where $f(q_{it} \mid k_{it}, \Phi, \eta_i, \delta_i, Z_{it}, X_j)$, $f(k_{it} \mid \tilde{\beta}_{it}, \Phi, \eta_i, \delta_i, \lambda_i, Z_{it}, X_j)$, and $f(\tilde{\beta}_{it} \mid \Phi, \eta_i, Z_{it})$ are the expressions derived in the appendix, (App-1), (App-2), and equation (21) in the main paper. Expressions $f(\eta_i \mid \mu_\eta, \sigma_\eta)$, $f(\delta_i \mid \mu_\delta, \sigma_\delta)$, and $f(\lambda_i \mid \mu_\lambda, \sigma_\lambda)$ correspond to the mixing distribution for the population parameters, as specified in the Model section. We choose diffuse prior distributions for all population parameters. We use a normal distribution with mean and standard deviation (0,100) for $\mu_\eta, \mu_\delta, \mu_\lambda$, and inverse-gamma with shape and scale parameters $(1, \sqrt{10})$ for $\sigma_\eta, \sigma_\delta, \sigma_\lambda$. We assume that $\Phi = \{b, \rho_1, \rho_2, \beta_0, a_1, a_2, \exp(r)\}$ follows a multivariate normal distribution with parameters $\mu_\Phi = [\varnothing_{n\Phi-1}, 3]$ and $\text{diag}(\Sigma_\Phi) = [100 \times \mathrm{I}_{n\Phi-1}, 1]$, where $n\Phi$ is the dimension of Φ, $\varnothing_{n\Phi-1}$ is a $1 \times n\Phi$ vector of zeros, and I_n is the identity matrix of dimensions $n \times n$. (The values of μ_Φ and Σ_Φ were chosen to ensure uninformative priors in the transformed space.) We draw recursively from the following posterior distributions:

1. (Gibbs) Parameters $\mu_\eta, \sigma_\eta, \mu_\delta, \sigma_\delta, \mu_\lambda, \sigma_\lambda$ are obtained by sampling from the following distributions:

- $f(\mu_\eta \mid \sigma_\eta^2, \eta_i) = \text{Normal}\left(\frac{\sum_{i=1}^{I}\eta_i}{I}, \left(\frac{1}{10000} + \frac{I}{\sigma_\eta^2}\right)^{-2}\right)$.

- $f(\sigma_\eta \mid \mu_\eta, \eta_i) = \text{Inverse Gamma}\left(1 + \frac{I}{2}, \sqrt{10} + \frac{\sum_{i=1}^{I}\left(\eta_i - \mu_\eta\right)^2}{2}\right)$.

We proceed similarly for parameters $\mu_\delta, \sigma_\delta, \mu_\lambda, \sigma_\lambda$.

2. (MH) Draws for Φ are obtained by sampling from

- $f(\Phi \mid \mu_\Phi, \Sigma_\Phi, \varpi_i, \beta_{it}, \text{data}) \propto \exp\left(.5\left(\Phi - \mu_\Phi\right)' \Sigma_\Phi^{-1}\left(\Phi - \mu_\Phi\right)\right) P(\text{data} \mid \varpi_i, \beta_{it}, \Phi)$

3. (MH) Draws for η_i are obtained by sampling from:

- $f(\eta_i \mid \mu_\eta, \sigma_\eta, \Phi, \delta_i, \gamma_i, \beta_{it}, \text{data}) \propto \exp\left(.5\frac{\left(\eta_i - \mu_\eta\right)^2}{\sigma_\eta^2}\right) P(\text{data} \mid \Phi, \varpi_i, \beta_{it})$

- We proceed similarly for λ_i.

4. (MH) Draws for β_{it} are obtained by sampling from:

- $f(\beta_{it} \mid \Phi, \delta_i, \gamma_i, \beta_{it}, \text{data}) \propto g\left(\beta_{it} \mid r\beta_0 + nr, \beta_0 + \sum_{t=1}^{n} s_{i\tau_t}\right) P(\text{data} \mid \Phi, \delta_i, \gamma_i, \beta_{it})$,

where $g\left(\beta_{it} \mid r\beta_0 + nr, \beta_0 + \sum_{t=1}^{n} s_{i\tau_t}\right)$ is the gamma pdf as derived in (21).

Since there is no closed-form expression for the posterior distributions of Φ and ϖ_i, we use a Gaussian random-walk Metropolis-Hasting algorithm to draw from these distributions. Following the Metropolis-Hasting procedure proposed by Atchade (2006), for each iteration, *s*, we draw a proposal vector of parameters $\zeta^{(s)}$ (either for α and ϖ_i):

$$\zeta^{(l)} \sim \text{Normal}\left(\zeta^{(l-1)}, \sigma^{(l-1)}, \Delta^{(l-1)}\right)$$

and then accept the vector using the Metropolis-Hastings acceptance ratio. The tuning parameters $\sigma^{(l-1)}$ and $\Delta^{(l-1)}$ are adapted in each iteration to get an acceptance rate of approximately 20%. We ran the simulation for 30,000 iterations. The first 20,000 iterations were used as a ``burn-in" period, and the last 10,000 iterations were used to estimate the conditional posterior distributions. Figure A9 and Figure A10 show the posterior draws obtained in the simulation.

Figure A9: Posterior draws for the population parameters (MH steps)

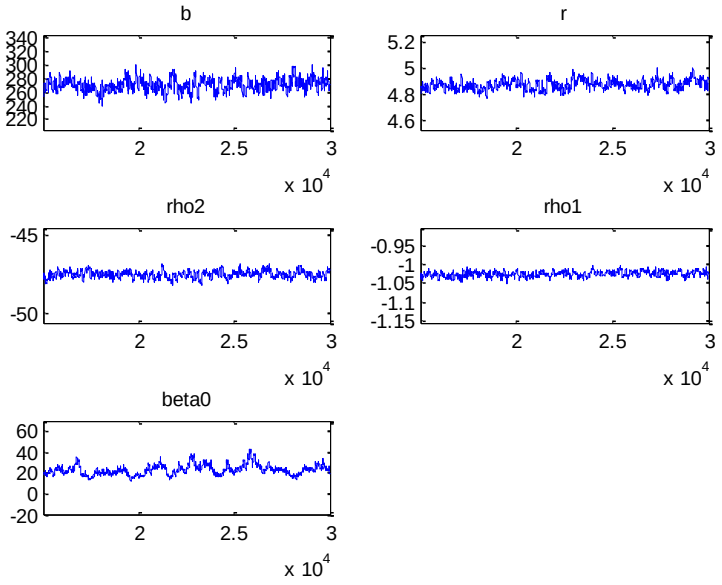
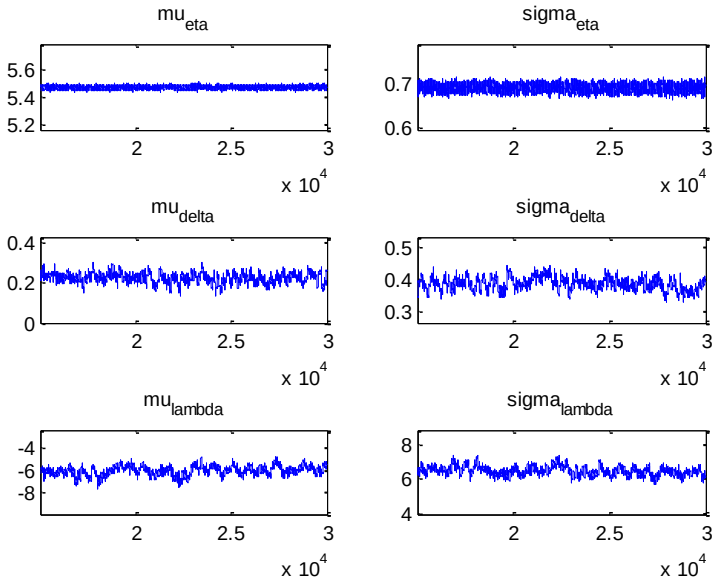


Figure A10: Posterior draws for mixing individual-level parameters (Gibbs)



FURTHER ROBUSTNESS CHECKS

Sensitivity analysis for the effect of switching costs on counterfactual analyses

Our econometric model assumes that customers' choice decisions are based on the next period only. This assumption does not affect the estimates of our main variable of interest, δ_i, but could potentially lead us to overestimate consumers' sensitivity to the switching fee, ρ_1. If this were the case, the effect of lowering the switching fee on provider revenues could be lower than what our results about recommendations to the firm suggest. We run a sensitivity analysis to measure whether the effect of reducing the switching fee, as presented in in the main manuscrip, is robust to lower levels of ρ_1. We reduce the estimate of ρ_1 by 5%, 10%, and 20%.

Figure A11, Figure A12 and Figure A13 replicate the results obtained in the main manuscript (see Figure 3 of the main manuscript) for lower levels of ρ_1. We find that the revenue impact from lowering the switching fee is very robust to lower levels of ρ_1. In an additional analysis, we similarly vary the level of the sensitivity to cost of switching to a different provider,

ρ_2. We find that while a lower sensitivity to cost of leaving the provider affects the level of provider revenues, it does not change the optimal level of the switching fee. Hence, we are confident that the assumption that customers make tariff choice decisions taking into account their usage in the next period only does not significantly bias our policy simulations.

Figure A11: Change in revenue due to reduction of the switching fee if ρ_1 is reduced by 5%

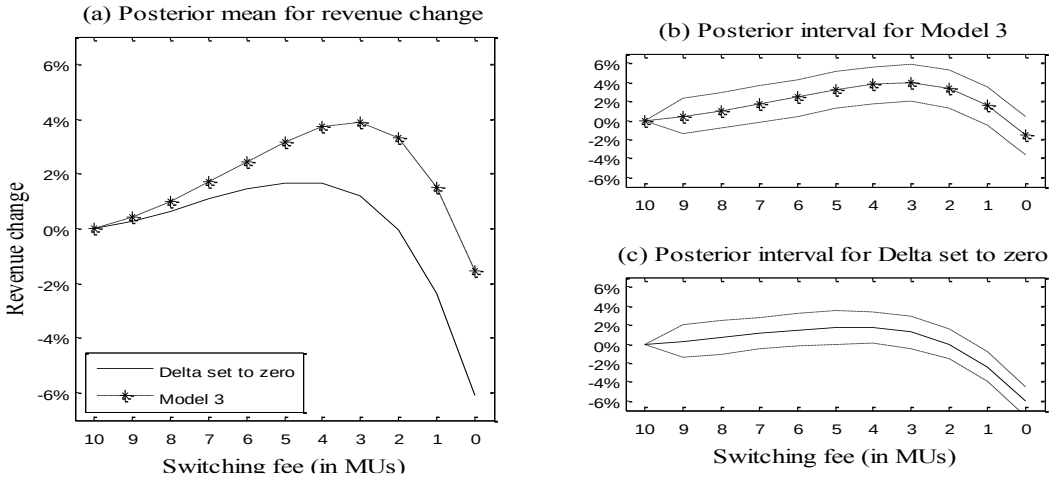


Figure A12: Change in revenue due to reduction of the switching fee if ρ_1 is reduced by 10%

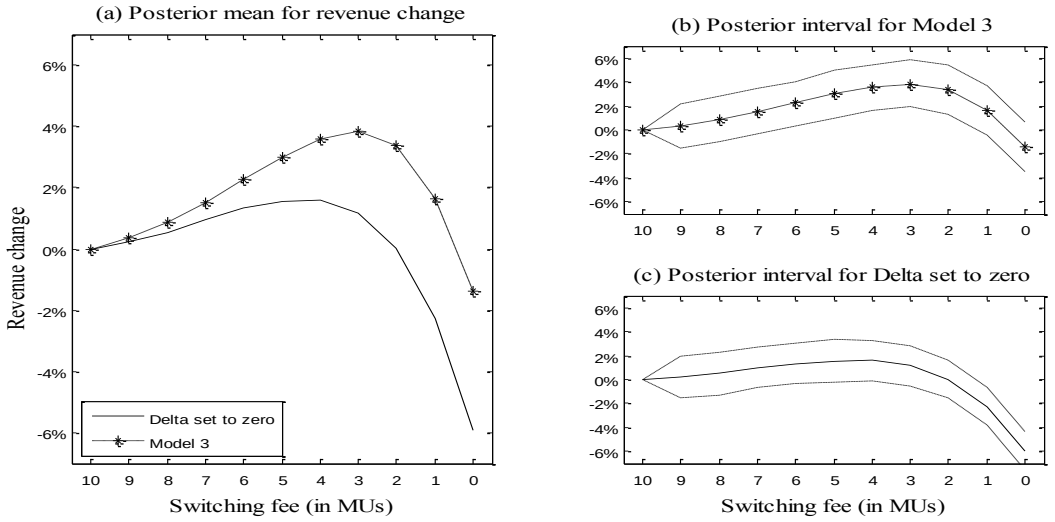
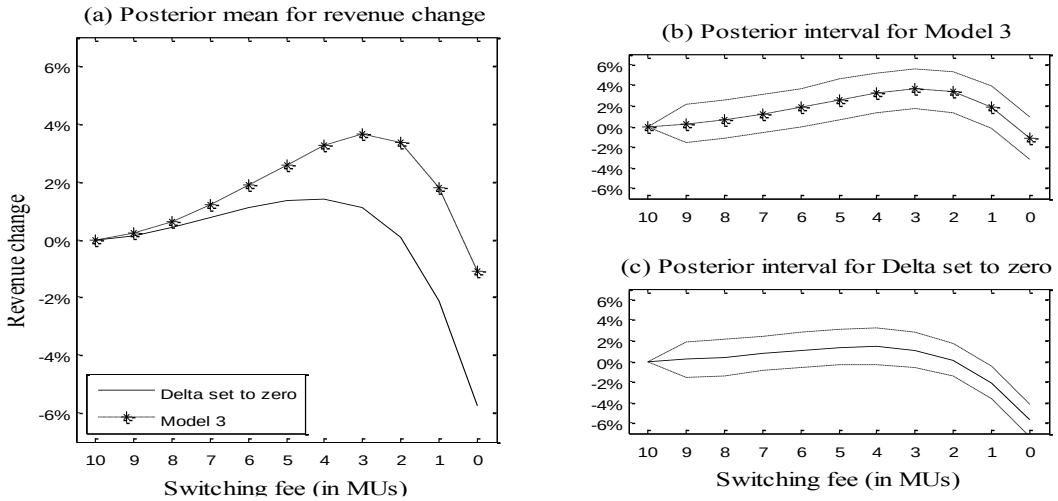


Figure A13: Change in revenue due to reduction of the switching fee if ρ_1 is reduced by 20%



Alternative model specification: Additional value for "free" minutes only

An alternative way to build our model would be to assume that three-part tariff customers assign greater value only to minutes strictly below the allowance, and not to all three-part tariff minutes. In our data, three-part tariff usage mostly lies beyond the allowance: 72% of three-part tariff observations exceed the usage allowance, by an average of 88.4%. As a consequence, a behavioral theory that limits the effect of free minutes to usage below the allowance seems, in principle, unable to explain the pattern in our data.

To further confirm this claim, we re-estimate Model 2 as presented in the main manuscript but allow the effect of free minutes, δ_i, to apply to minutes within the allowance only. We find that such a model does not reflect the phenomenon we observe well. First, the fit is worse than that of Models 2 and 3 (Model section of the main manuscript) that assume that the additional valuation δ_i applies to all three-part tariff minutes. The MSE of the alternative model is 55.08 versus a MSE of 46.33 in Model 2 and 45.54 in Model 3. In the alternative model we obtain a MAPE of 75.74 versus a MAPE of 72.4 in Model 2 and of 71.87 in Model 3. Second, we obtain a negative

posterior mean of the variable relating to the value of free minutes. This estimate is negative because in our sample customers generally consume above the allowance. As a consequence, a model that only estimates δ_i from minutes within the allowance would overestimate the satiation level for customers who switch to a three-part tariff and often consume above the allowance (i.e., the majority of our three-part tariff customers). Then, in the periods in which these customers consume within the allowance, δ_i needs to be negative to compensate for the overestimation of their satiation level. A negative delta cannot explain the usage increase observed in the data and documented in the main manuscript, and it is not consistent with previous literature indicating that "free" would lead to increased valuation of the good. We conclude that this model specification is not a good representation of the phenomenon we observe.

Table A8: Posterior distribution of parameter estimates for model where δ_i applies to free minutes within the allowance only

	Model 2 "free" minutes only		
	Mean	95% Interval	
Demand intercept			
Mean, μ_η	5.520	5.501	5.538
Std. dev., σ_η	0.690	0.679	0.701
Demand slope, *b*	240.356	230.992	250.097
Variance of usage shock, *1/r*	0.216	0.214	0.218
Valuation of free units			
Mean, μ_δ	-0.301	-0.364	-0.234
Std. dev., σ_δ	0.478	0.433	0.536
Preferences in tariff choice, ζ_{ijt}			
SC bw. tariffs, ρ_1	-1.025	-1.035	-1.014
SC to other provider, ρ_2	-47.792	-48.164	-47.425
Preference for the three-part tariff			
Mean, μ_λ	-4.749	-5.226	-4.252
Std. dev., σ_λ	5.556	5.250	5.858
Log Marginal Density	-394749		
MSE ('000)	55.08		
MAPE	75.74		

N=5,831 customers, 63,449 usage and 63,616 choice observations
Demographic shifters of the demand slope included but not reported for readability.

REFERENCES

Atchade, Yves F. (2006), “An adaptive version for the Metropolis adjusted Langevin algorithm with a truncated drift,” Methodology and Computing in Applied Probability, 8, 235-254.

Blundell, Richard, and Stephen Bond. (1998). "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models," Journal of Econometrics 87, 115-143.

Roodman David (2009) "How to do xtabond2: An introduction to difference and system GMM in Stata," Stata J.,9, 86-136