Is Tipping Economically Relevant? Evidence from the Beauty Industry

Jacob Kohlhepp[*] Danny Ober-Reynolds[†]

April 26, 2021

Abstract

In this paper we demonstrate tips are sensitive to service quality even when future interaction is unlikely. Using a novel data set covering 150,000 hair salon appointments where customers can be observed over time, we are able to exploit variation in service quality and exogenous separation rates. This allows us to separate the dynamic and direct effect of service quality on tips. We show that an important part of tipping behavior is a social norm for quality: clients tip based on perceived quality even when they do not expect to see the stylist again. At the same time, dynamic concerns make tips more sensitive to quality. We show in a stylized dynamic model how such a social norm for quality can support greater effort provision in equilibrium. Our results support the view of tipping as a social norm which encourages cooperation.

Keywords: tipping, haircuts, social norms, dynamic behavior

JEL Codes: J32, D91

[*] Department of Economics, UCLA. Contact: jkohlhepp@ucla.edu.

[†] Department of Economics, UCLA. Contact: doberreynolds@g.ucla.edu.

1 Introduction

Tipping is a massive economic phenomenon in the United States, representing an important part of compensation for many American workers. While tipping is a common practice, it is not clear whether it is economically meaningful. Are tips merely transfers? Or do they sustain higher levels of service quality when future interactions are uncertain?

Consider a simple static model. A customers pays for a service. The stylist then performs the service and can exert costly effort to improve quality. The customer can leave a tip after observing quality. Absent any sort of norm, the players are stuck in an inefficient equilibrium where the customer leaves no tip for all levels of service quality and the stylist anticipates this and exerts no effort.

In this situation, the issue is the inability of the client to commit to a tip based on quality. If the client could commit, greater quality and social surplus could be supported in equilibrium. Social norms can provide this commitment: even when a client knows they are never going to come back social costs can compel them to leave a tip that matches the quality of the service. This is all well and good in theory. We seek to understand whether this social norm exists in practice.

In this paper, we show tips are sensitive to quality even when future interaction is unlikely. That is, we provide evidence for a *social norm for quality.* Using a novel data set covering 150,000 hair salon appointments where customers can be observed over time, we are able to infer hair stylist quality. This allows us to use a mediation model to separate the impact of dynamics and quality on tips. We estimate a service at the 75th percentile of quality receives an expected tip which is 1 percentage point higher than a service at the 50th percentile *even when a client is not planning on returning.* Dynamics make tips more quality sensitive: a service at the 75th percentile receives a tip that is more than 2 percentage points higher than a service at the 50th percentile.

We contribute to the literature on tipping in the United States. The literature so far has focused on the case of restaurants and more recently, Uber drivers. While these are interesting, both are situations where concerns about the future are second-order. People do not usually match with the same Uber driver again, and restaurant workers have high turnover rates. People often try to return to the same stylist and the same salon once they have found someone they like. Also, measuring the quality in the restaurant case is quite difficult. Indeed, most studies so far have used self-reported surveys. We develop a methodology to infer quality using observed behavior.

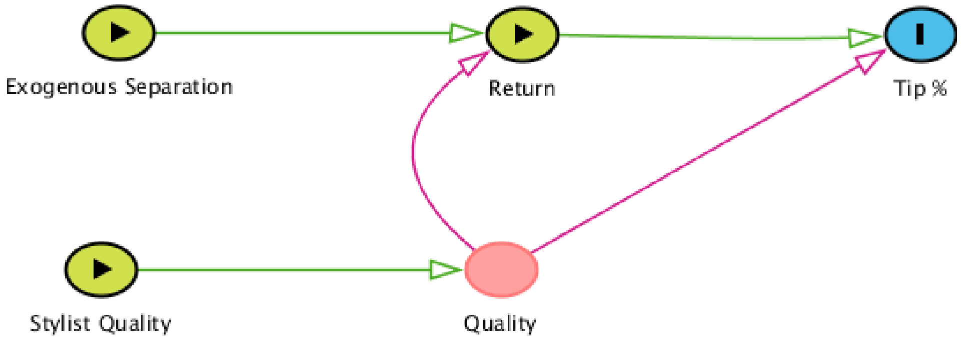


Figure 1: Representation of Variable Relationships

2 Conceptual Framework

Index clients by i and stylists by j. We will focus on only the first observe ditneraction of each pair, so we can drop time subscripts. Stylists are characterized by their time-invariant general quality a_j. When a client and stylist match, the quality of a haircut $(q_{i,j})$ is given by the sum of general quality plus an i.i.d. match component denoted $m_{i,j}$.

After the haircut, the client decides whether they will return $(r_{i,j})$. This decision is based on quality of the haircut as well as exogenous factors which we denote $e_{i,j}$. At the same time the client can choose a tip which we denote $b_{i,j}$. The tip will be a function of quality as well as whether the client is returning. We can characterize the structural equations of this model as:[1]

$$q_{i,j} = a_j + m_{i,j}$$

$$r_{i,j} = g(q_{i,j}, e_{i,j})$$

$$b_{i,j} = f(q_{i,j}, r_{i,j}, u_{i,j})$$

We depict these relationships through a directed acyclic graph in Figure 1. The figure suggests how we will eventually identify the indirect and direct effects of quality on tip percentage.

Using this notation we can formalize the effects we are interested in. First, we want to understand if tips are sensitive to quality even when the client is not returning. We call this **behavioral quality sensitivity** because it is only rational if we incorporate some sort of psychic cost. In order to understand how the norm

[1] This is a mediation model in the style of Pinto (2019), because there is a direct effect of quality on tips and a **mediated effect** through the return decision. The main difference is that quality (q) is not observed, so we cannot apply their results directly.

depends on quality we can examine the following expression.

$$E_u[f(q,0,u)|q=x]$$

This should be upward sloping if there is behavioral quality sensitivity. We also want to understand the dynamic effect. We can measure this as the difference between the tip of a returner and a non-returner:

$$E_{q,u}[f(q,1,u)-f(q,0,u)]$$

Finally, we can study whether tips are more sensitive to quality when someone is returning. This can be measured as:

$$\frac{\partial}{\partial x}E_u[f(q,1,u)-f(q,0,u)|q=x]$$

Although we do not formally model effort choice, this is without loss. To see why, note that any stylist making an effort choice must do so before knowing how someone will tip. That is, they must choose effort based on a_j and equilibrium beliefs about the sensitivity of the tip to quality. So whatever each stylist's effort choice it will be the same for all clients (unless we consider additional client heterogeneity). Effort will enter q_j as part of a_j. Since we do not constrain a_j in any meaningful way, this does not cause any practical issues. Thus we can think of a_j as being the total effect of equilibrium effort choice and general skill.

Now we specify functions. First the return decision:

$$g(q,e)=\mathbb{I}\{q\geq 0\}e$$

So e is the probability of returning conditional on quality being satisfactory. We assume that e is distributed i.i.d. across all clients within salon. We denote μ_k^e as the mean of e for clients at salon k. For the tipping decision we specify:

$$f(q,r,u)=\alpha_1 q+\alpha_2 r+\alpha_3 q\cdot r+u$$

That is we assume linearity with an interaction between return belief and quality. u can be thought of as a generosity shock. Putting this all together gives:

$$q_{i,j}=a_j+m_{i,j}$$

$$r_{i,j}=\mathbb{I}\{q_{i,j}\geq 0\}e_{i,j}$$

$$b_{i,j} = \alpha_0 + \alpha_1 q_{i,j} + \alpha_2 r_{i,j} + \alpha_3 q_{i,j} \cdot r_{i,j} + u_{i,j}$$

Under this specification, the three effects defined earlier have simple expressions. Behavioral quality sensitivity is given by:

$$E_u[f(q,0,u)|q=x] = \alpha_0 + \alpha_1 x$$

The average dynamic effect is given by:

$$E_{q,u}[f(q,1,u) - f(q,0,u)] = \alpha_2 + \alpha_3 E[q]$$

And the increased quality sensitivity due to dynamic concerns is given by:

$$\frac{\partial}{\partial x} E_u[f(q,1,u) - f(q,0,u)|q=x] = \alpha_3$$

3 Model Comments

Our model is not meant to be a full specification of a dynamic game. However, it can be thought of as the reduced form of a family of dynamic games where the client has a tipping strategy based on quality and whether there will be a future and the stylist can choose effort. The coefficient on quality represents the tipping strategy of the client in the first period when future interaction is unlikely. The coefficient on the return indicator and the interaction represents the adjustments the client makes to their tipping strategy given a future value function. Such a value function will depend mainly on the quality of the service. Thus, while we do not specify a formal game, our framework allows us to explore the strategies that result from whatever game is being played in equilibrium.

One caveat worth mentioning is that our specification assumes the client knows fully whether or not he/she will return. In most games there is some form of exogenous separation which cannot be fully anticipated. We do not include this in the statistical model because it adds another layer of complexity to estimation. We instead argue that our model is a good first-order approximation, because clients usually have close to full knowledge about whether or not they are coming back when they leave a tip.

4 Data

We use proprietary data from Boulevard, a salon transaction software company. The company sells a software product which allows salons to track appointments, inventory and receipts.

5 Estimation

Suppose for now that stylist time-invariant quality, a_j, is fully observed. We re-write the tip equation as:

$$b_{i,j} = \alpha_0 + \alpha_1 a_j + \alpha_2 r_{i,j} + \alpha_3 a_j r^*_{i,j} + u_{i,j} + m_{i,j}(\alpha_1 + \alpha_3 r_{i,j})$$

Because the return decision depends on $m_{i,j}$ this cannot be estimated with OLS. Define $X_{i,j} = (1, a_j, r^*_{i,j}, a_j \cdot r_{i,j})$. Consider estimating for salon k. Then we have:

$$E[(b_{i,j} - X'_{i,j}\alpha)X'_{i,j}] = \begin{bmatrix} E[m_{i,j}(\alpha_1 + \alpha_3 r_{i,j})] \\ E[a_j m_{i,j}(\alpha_1 + \alpha_3 r_{i,j})] \\ E[m_{i,j}(\alpha_1 r_{i,j} + \alpha_3 r^2_{i,j})] \\ E[a_j m_{i,j}(\alpha_1 r_{i,j} + \alpha_3 r^2_{i,j})] \end{bmatrix} = \mu^e_j \begin{bmatrix} \alpha_3 E[\phi(-a_j)] \\ \alpha_3 E[a_j \phi(-a_j)] \\ \alpha_1 E[\phi(-a_j)] + \alpha_3 E[\phi(-a_j)] \\ \alpha_1 E[a_j \phi(-a_j)] + \alpha_3 E[a_j \phi(-a_j)] \end{bmatrix}$$

We have 4 parameters and 4 moments, so $\alpha_0, \alpha_1, \alpha_3, \alpha_4$ are constructively identified (and therefore estimable using GMM) if a_j and μ^e_k are known. However they are not observed or known. μ^e_j is the exogenous separation rate given that quality is sufficiently high. We assume that quality is fully learned after the first visit.[2] This implies that the probability of returning a 3rd time conditional on coming twice is equal to e. Thus we can use the following sample estimator:

$$\hat{\mu}^e_j = \frac{\sum_i r^3_{i,j}}{\sum_i r^2_{i,j}}$$

where r^2, r^3 are indicators for whether a client returned a 2nd or third time. To estimate a_j, notice that the average probability of staying at stylist j yields information about quality. Specifically:

$$E[r_{i,j}] = \Phi(a_j)\mu^e_j$$

Since we have already estimated μ^e_j our estimator is given by:

$$\hat{a}_j = \Phi^{-1}\left(\frac{I^{-1}\sum_i r_{i,j}}{\hat{\mu}^e_j}\right)$$

We can now estimate a_j, μ^e_j in a first stage then plug-in the estimates and perform GMM in a second stage. Inference is complicated by the two stages. We therefore obtain standard errors by block bootstrapping the

[2] We can relax the learning assumption as much as we want by using progressively higher moments at the cost of statistical power. We do this in the Appendix as a robustness check.

entire process.[3]

6 Results

The estimated coefficients, along with the 3 main effects are presented in Table 1.

Table 1: Tipping Model Estimates

	Percent Tip			
	(1)	(2)	(3)	(4)
Quality	1.129**	1.443***	0.214	$-$0.051
	(0.512)	(0.470)	(0.521)	(0.555)
Return	$-$1.968***	$-$2.186***	$-$0.560	$-$0.311
	(0.660)	(0.693)	(0.680)	(0.722)
Quality x Return	1.520***	1.430**	0.931**	0.891*
	(0.550)	(0.560)	(0.453)	(0.488)
Fixed Effects:	State	City	Business	Salon
Haircuts	45509	45509	45509	45509
Stylists	406	406	406	406
Firms	34	34	34	34
Salons	48	48	48	48

Bootstrapped Standard Errors: *p<0.1; **p<0.05; ***p<0.01

Note: Standard errors are from a block bootstrap with resampling at the stylist level (100 replications).

We estimate the tipping equation using four specifications. Each specification includes progressively more fine-grained fixed effects. Our preferred specification is specification 2, which includes city fixed effects. State fixed effects are too broad: tipping norms are probably established more locally. Business fixed effects are two narrow: they remove much of the identifying variation in stylist quality. We believe this is why the coefficients change dramatically from column two to three.

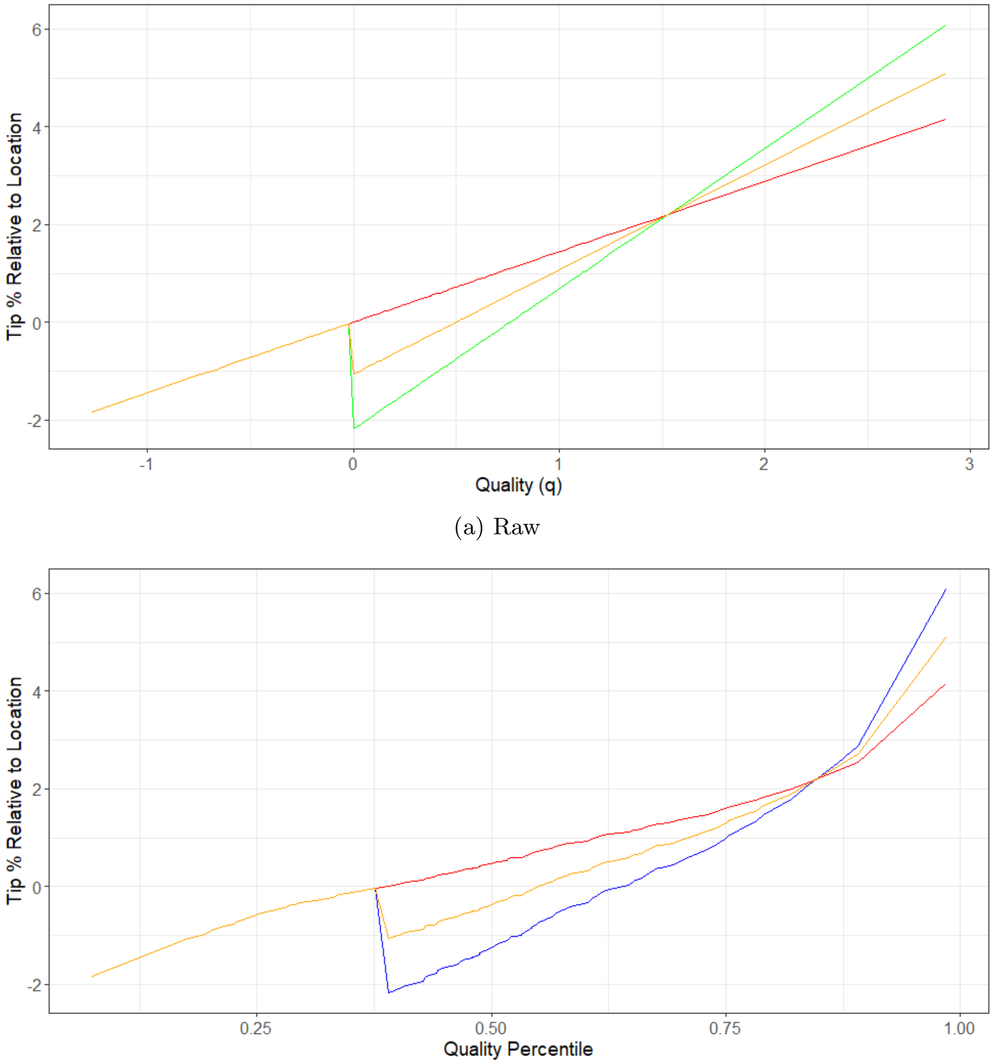
Under specification 2, both quality and the return-quality interaction have positive and statistically significant coefficients. When a client is not returning, a one unit increase in quality increases expected tip by 1.4 percentage points. We interpret this as evidence of a behavioral norm for quality, because clients are tipping in a way that cannot be supported in a sub-game perfect Nash Equilibrium. When a client is returning, the effect doubles. A one unit increase in quality increases expected tip by 2.9 percentage points. We take this as evidence that dynamic concerns magnify quality sensitivity. Overall, these estimates make

[3] Blocks are at the stylist level.

the case that the tipping norm essentially stands in for dynamic concerns. They make a client act as if there is a future even when the client knows they are not coming back.

We visualize the connection between quality and tip percentage in Figure 2.

Figure 2: Expected Tip Conditional on Quality



(a) Raw

(b) percentile

To understand Figure 3 Panel A, recall that a client may not return for two reasons. Either service quality is too low ($q < 0$) or there is exogenous separation ($e = 0$). The three lines represent the following cases:

1. **Green:** Expected tip conditional on the client coming back if quality is good enough ($e = 1$).

2. **Red:** Expected tip conditional on the client not coming back even if quality is good enough ($e = 0$).

3. **Orange:** Unconditional expected tip.

Notice that the orange line lies between the red and green lines. This is because the unconditional expected tip is the weighted average of the other two functions, where the weights are determined by the exogenous separation rate μ_j^e.[4] The figure illustrates several striking features of our results. First, expected tip conditional on quality drops at 0 when there is not exogenous separation. This is why both the green and orange lines exhibit a discontinuity at 0. Second, this drop means that expected tip is lower for just-good-enough quality when the client is going to return then when he/she is not going to return. In the figure, this is represented by the green line dipping below the red line at 0. Third, tips are more sensitive to quality when the client is not going to exogenously separate. In the figure, this is represented by the steeper slope of the green line relative to the red line.

We also present a similar graph with a more interpretable x-axis in Panel B. We transform quality into quality percentile using the estimates of stylist quality combined with our assumption that match-specific quality is independent standard normal.[5] Quality percentile represents the rank of a haircut in the quality distribution, with 1 being a haircut of the best possible quality and 0 being a haircut of the worst possible quality. We see that the quality cutoff occurs around the 38th percentile; that is about 38 percent of the time clients separate because quality is not high enough.

Finally, we present the unconditional tip percentage (orange function) with confidence bands on its own.[6] We think it is worthwhile paying closer attention to this graph, because it represents the total effect of service quality on tips. Notice that the drop at the quality threshold is less pronounced, and as a result expected tip is close to monotonic in quality. We present this graph to reiterate one point: clients generally tip more for higher quality.

[4] In the figure, we use the sample average μ^e over all locations, but the same graph can be produced for each of the μ_j^e estimates.

[5] The CDF of quality under these assumptions is given by $F_q(q) = \int F_a(q-m)d\Phi(m)$. We derive the empirical CDF of stylist quality $\hat{F}_a$ and then use quadrature to estimate the integral.

[6] Graphing all 3 lines together with confidence bands in visually overwhelming.

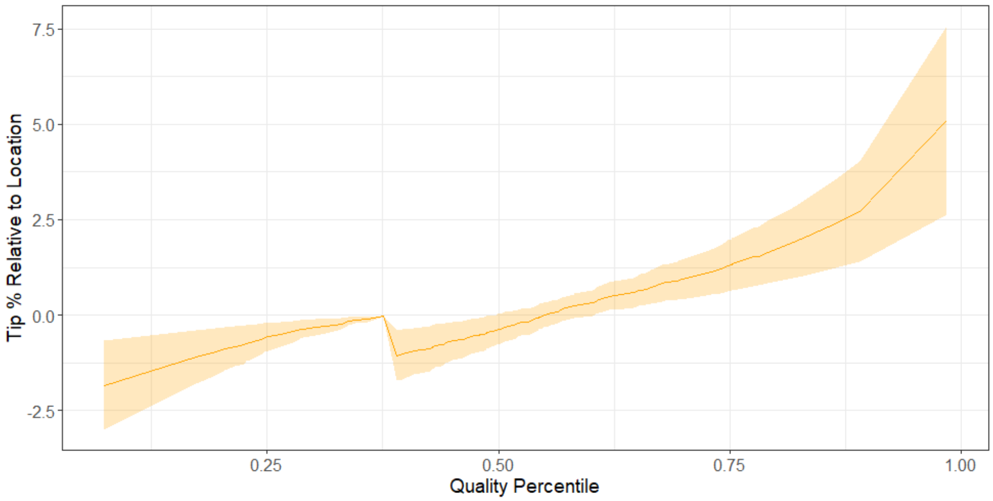


Figure 3: Total Effect of Service Quality on Tip percentage

7 Equilibrium Effort Provision: Under Construction

In this section we extend our model to think about equilibrium effort provision. Suppose stylists control effort, z which in turn impacts their stylist specific quality. Thus we have that $a'(z) > 0$.

8 Taxation of Tips: Under Construction

So far we have established that tipping is sensitive to quality. We have also argued that if quality depends on stylist effort, this also means that a tipping social; norm supports a more efficient level of equilibrium effort. This result holds importance for the question of taxing tips. Consider a government which can collect an ad valorem tax that is $1 - \tau$ percent of the the tip.[7]

Suppose we abstract from the return decision and instead assume that $r^* = 0$. Then the derivative of the tip with respect to a_j is α_1. Consider a simple example where cost is quadratic and given by $e^2/2$. Putting this together we have that equilibrium effort is given by:

$$c'(e_j^*) = \frac{\partial q}{\partial e} \leftrightarrow e_j^* = \alpha_1 \tau \frac{\partial a_j}{\partial e^*}$$

where the unevaluated derivative is the impact of effort on a_j. If we assume this has a linear impact, we can replace this with a constant like so:

[7] Currently taxes are technically income, so they are taxed at an effectively ad valorem rate, but there is a little enforcement. So this could also be thinking about the effect of stronger enforcement of existing tax law.

$$e_j^* = \alpha_1 \tau \Delta_j$$

The deadweight loss resulting from taxation is then given by (not accounting for the revenue generated):

$$\alpha_1 \Delta_j (1 - \tau) - (\alpha_1 \Delta_j)^2 (1 - \tau^2)$$

Note that this is increasing for all $\alpha_1 < [\Delta_j (1 + \tau)]^{-1}$, and decreasing otherwise.