

**The Effects of Tip Recommendations
on Customer Tipping, Satisfaction, Re-patronage, and Spending**

Damon Alexander

ex Washio

Christopher Boone

School of Hotel Administration

Cornell University

Michael Lynn*

School of Hotel Administration

Cornell University

* Order of authorship is intentionally alphabetical. Address correspondence regarding this paper to Mike Lynn, 552 Statler Hall, Cornell University, Ithaca, NY 1483-6902, WML3@cornell.edu, (607) 255-8271. The authors would like to thank Jordan Metzner, CEO and co-founder of Washio, for conducting the experiment reported here and for allowing us to analyze the data and publicly report the results in this paper.

ABSTRACT

A field experiment involving 94,571 orders from 24,637 customers of an app-based laundry pick-up, cleaning and delivery service examined the effects of various, randomly assigned tip recommendations on consumers' tip amounts, satisfaction ratings, frequency of return, and bill size. We find that tip recommendations affect tip amounts, but not customer satisfaction, patronage frequency, or bill-size, which implies that neither the processes underlying the tip recommendation effects on tipping nor consumer tipping itself affect these other consumer outcomes. From a practical perspective, these results and conclusions inform efforts to increase or decrease tipping. Recommending larger tip amounts, at least within the \$2 to \$10 or 5% to 25% ranges studied here, appears to be a safe means of increasing the amounts customers leave. More generally, altering customers' tipping behavior will not itself adversely affect those customers' subsequent satisfaction, re-patronage, or spending as long as the means used to alter tipping do not directly affect these other outcomes.

The Effects of Tip Recommendations on Customer Tipping, Satisfaction, Re-patronage and Spending

1. Introduction

Every day, millions of consumers around the world give voluntary payments of money (called “tips,” “propinas,” and “trinkgelds,” among other names) to service providers who have already served them. Among those service workers commonly receiving tips are baristas, bartenders, casino dealers, concierges, deliverymen, doormen, hair cutters, hotel maids, masseurs, porters, sommeliers, street musicians, taxi drivers, tour guides, and waiters (Lynn, 2016). Precise measures of the total amount tipped worldwide do not exist, but estimates place the amount tipped to restaurant workers in the U.S. alone at over \$45 billion a year, so the former number must be enormous (Azar, 2011). To service workers, these tips represent a source of income, of job motivation, and of subordination to consumers (Namasivayam and Upneja, 2007; Shamir 1983; Shy, 2015). To consumers, these voluntary gifts represent an avoidable cost, a normative obligation, an expression of satisfaction, an incentive for future service, and an attempt to help service workers (Becker, Bradley and Zantow, 2012; Lynn, 2015b). To service firms permitting tipping of their employees, these consumer payments represent a form of buyer-monitoring, employee compensation, pay-what-you-want pricing, price-discrimination, and price-partitioning (Azar, 2011; Lynn and Withiam, 2008; Schwartz, 1997).

The complexity and theoretical richness of this behavior has attracted the interest of scholars in anthropology, economics, hospitality management, human resources management, marketing, social psychology, sociology, and tourism management (see Azar 2007; Lynn 2006, 2015b for reviews). Research on this topic has examined both ways to increase the tips consumers leave (Lynn, 2018; Seiter, Brownlee and Sanders, 2011) and the organizational consequences of permitting and encouraging tipping (Azar, 20011; Lynn, 2017). We contribute

to both streams of research by investigating the effects of tip recommendations on customers' tipping, satisfaction, re-patronage, and spending.

Our data come from a now defunct app-based laundry pick-up, cleaning, and delivery service. The app that customers used to place and pay for orders also allowed them to add a tip for the delivery drivers to their bills. Specifically, the app presented customers with three suggested tip amounts (each of which could be selected with one click), an option to leave a custom tip amount, and a one-click option not to tip (see Figure 1). The company presented different customers with different, randomly selected sets of suggested tip amounts, but presented the same customer with the same suggestions each time he or she used the app. The different sets of suggested tip amounts presented to different customers allowed us to examine the effects of recommending (i) large vs. small tip amounts, (ii) percentage vs. dollar tip amounts, (iii) a broad vs. narrow range of tip amounts (holding the average suggested tip constant), and (iv) round vs. non-round dollar tip amounts (\$x.00 vs. \$x.99).

We found that, in the context of providing several different tip recommendations, the provision of larger tip recommendations increased the size of tips that were left as well as overall tip revenues - even while decreasing the likelihood of tipping when the recommendations exceeded familiar and normative amounts. Larger tip recommendations had no corresponding effects on customer satisfaction, re-patronage frequency, or bill sizes. These findings suggest that larger tip recommendations' do not evoke a negative affective reaction that jeopardizes future patronage - even from those people who are dissuaded from tipping by the larger tip recommendations. Thus, firms can increase employee income by recommending larger tips (within the range studied here) without fear of negative customer reactions. The findings also suggest that tipping behavior itself has little positive or negative effect on the tippers' subsequent

attitudes or behaviors, so service firms can set tipping policies with little regard for such effects. Additional findings indicated that: (i) recommending equivalent tips in percentage instead of dollar formats increased tip amounts when some of the recommended percentages were sub-normative, (ii) increasing the range of suggested tips increased the likelihood of choosing one of the recommended amounts, but did not affect the amount tipped on average, and (iii) recommending equivalent round and non-round tip amounts (differing in size by only a penny, but with different left-most as well as right-most digits) did not produce different levels of tipping.

2. Related Literature

Our study contributes to knowledge about ways to increase the tips consumers leave as well as about the organizational consequences of permitting and encouraging tipping. We also draw from, and build on, a related literature concerned with the effects of different appeals for charitable donations.

2.1. Effects of increasing recommended tip sizes

The tip recommendations in this study included several sets of recommendations that differed in magnitude of the recommended amounts but were otherwise similar to one another. Our examination of the effects of these differences in recommended tip size most directly builds on Haggag and Paci's (2014) study of default tips in taxicabs. Using a large database of NYC taxicab rides, they tested (i) the effects of one firm's use of different default tip amounts and formats (\$ vs. %) for fares above and below \$15, and (ii) the effects of differences between two firms' percentage default tip amounts on fares above \$15, to assess the impact of different default sizes. Across both specifications, they found that larger default tip options were associated with a lower likelihood of tipping, but with larger tip amounts from those who did tip.

The empirical setting presents some challenges: namely, the impact of default tip size is confounded with tip format (\$ vs. %) in the first specification, and the effects in the second specification may be confounded by other differences between the companies. Nevertheless, the consistency of their results across different specifications supports the idea that it is the size of the default options that is driving the results.¹ Such an effect is plausible because suggested or default options convey information about the expected or typical contribution, which is likely to guide the behavior of those willing and able to conform with the implied norm, but may offend or discourage those unwilling or unable to do so (DeBruyn and Prokopec, 2013; Haggag and Paci, 2014). Our study conceptually replicates these findings in a natural field experiment with naïve subjects randomly assigned to un-confounded tip recommendation conditions, so it provides strong causal inferences about the effects of recommending larger tips in at least one naturally occurring setting (Al-Ubaydli and List, 2013).

Moreover, our dataset follows individual customers over time and includes their satisfaction ratings, which permits us to test the effects of the tip recommendations on outcomes of interest that were unavailable to Haggag and Paci (2014) – namely customer satisfaction, re-patronage frequency and subsequent spending. Recommending larger tips might affect these other outcomes either directly or indirectly through their effects on tipping. The direct effects (if any) are likely to be negative because large tip recommendations are more likely than small ones

¹ Further supporting this conclusion is the fact that similar effects have been observed in charitable donation contexts, where larger asks have been found to: (i) decrease the likelihood that donations will be made, (ii) increase the size of those donations that are made, and (iii) have variable effects on total revenue generated (Desmet, 1999; DeBruyn and Prokopec, 2013; Doob and McLaughlin, 1989; Schibrowsky and Peltier, 1995; Weyant and Smith, 1987).

to be considered unreasonably high or construed as threats to freedom of choice, either of which could generate negative affect (Carr, 2007; Clee and Wicklund, 1980; Fitzsimmons and Lehmann, 2004). As noted by Haggag and Paci (2014), such negative cognitive and affective reactions might also explain why larger asks decrease the likelihood of tipping.

The indirect effects of larger tip recommendations on customer satisfaction, re-patronage and spending through their effects on tipping are potentially more complex and could be negative, positive, or both. First, the larger tips evoked by larger tip recommendations may be viewed or felt by consumers as cost increases, which could lower customers' perceptions of value, satisfaction, re-patronage frequency and subsequent spending consistent with the law of demand (Varki and Colgate, 2001).

Second, tip recommendations that evoke larger tips could increase customers' satisfaction, re-patronage and spending because tipping activates self-perception, self-justification, or warm-glow processes. Tipping is a voluntary payment for services rendered as well as a pro-social behavior. Social psychological theory and research suggests that voluntary behavior often affects attitudes either because people infer their attitudes from their behavior (Fazio, 1987; Robins and John, 1997) or change their attitudes to justify, or be consistent with, their behavior (Aronson, 1969; Gawronski, 2012). Furthermore, marketing researchers have found that inducing customers to engage in pro-social behavior evokes positive feelings (called a "warm glow") that enhance customers' evaluations of service and retail experiences as well as their intentions to re-patronize the firms that provide those experiences (Giebelhausen and Chun, 2017; Giebelhausen, Chun, Cronin and Hult, 2016; Giebelhausen, Lawrence, Chun and Hsu, 2017). Thus, self-perception, self-justification, and warm glow processes may lead consumers to like, value and use a service more the more they tip for it.

Third, the opposing effects of larger tip recommendations on tipping likelihood and tip size could decrease non-tippers' satisfaction, re-patronage, and spending while simultaneously increasing tippers' satisfaction, re-patronage and spending. The positive effect on tippers' satisfaction, re-patronage, and spending of increasing the amounts they tip has already been discussed. Those same self-perception/justification and warm-glow processes could decrease non-tippers' satisfaction, re-patronage and spending. The opposing nature of these effects on tippers and non-tippers makes it hard to predict the direction of the average effect on the sample as a whole. However, if larger tip recommendations do have opposing effects on the satisfaction, re-patronage and spending of tippers and non-tippers then they should increase the variance in these other outcome variables. We test these competing expectations about the effects of larger tip recommendations, and of tipping itself, on customer satisfaction, re-patronage and spending for the first time.

2.2. Effects of recommending percentage vs. dollar tip amounts

The tip recommendations in this study included a set of small percentage tip amounts (5%, 10%, 15%) and a set of small dollar tip amounts (\$2, \$4, \$6) that were arithmetically identical when bill size was \$40. It also included a set of larger percentage tip amounts (10%, 15%, 20%) and a set of larger dollar tip amounts (\$4, \$6, \$8) that were arithmetically identical when bill size was \$40. Since the vast majority of bill sizes in our dataset exceeded \$40, the percentage recommendations were generally larger than the dollar recommendations and, given the effects of request magnitude discussed previously, should lead to larger average tips but a lower likelihood of tipping than do the dollar recommendations. However, when bill size was less than \$40, the dollar recommendations were larger than the percentage ones, so their effects on tipping should be reversed. More uncertain, but potentially interesting, is what happens when

bill size is around \$40 and the request magnitudes of these two types of recommendations are comparable. We tested all of these effects below.

2.3. Effects of recommending a larger range of tips

Among the tip recommendations in our study are two sets of dollar tip recommendations (\$2, \$4, \$6 and \$3, \$4, \$5) and two sets of percentage tip recommendations (10%, 15%, 20% and 12%, 15%, 18%) that had the same means but different ranges of recommended amounts. Ours is the first study that we know of to manipulate the range of several recommended tips or donations independently of the mean of those recommendations, so there are no existing findings on which to base expectations about the effects of this manipulation. However, DeBruyen and Prokopec (2013) found that when charitable appeals recommended several donation amounts, the size of the lowest recommendation had a larger effect on donations than did the sizes of the other recommendations. Furthermore, other researchers have found that legitimizing small contributions to charity increases the likelihood of giving without reducing the average gift size (Cialdini and Schroeder, 1976; Weyant and Smith, 1987) and that increasing the size of requested donations increases the amounts given (Doob and McLaughlin, 1989). These findings suggest that increasing the range of recommended tip amounts may increase the proportion of people leaving tips by legitimating smaller tip amounts as well as increase the size of those tips left by simultaneously asking for more. We empirically test this possibility for the first time.

2.4. Effects of recommending round vs. non-round tip amounts

The tip recommendations in this study also include one set of round dollar recommendations (\$4, \$6, \$8) and another set of non-round recommendations that were a penny less than the round ones (\$3.99, \$5.99, \$7.99). Round numbers are easier to process than non-round ones (Estelami, 1999) and round vs. non-round pricing has been shown to have numerous

effects on attitudes, beliefs and behavior. For example, consumers perceive round prices as more convenient than non-round ones (Wieseke, Kolberg and Shons, 2016), like round prices more than non-round ones (Guido and Peluso, 2004), choose round prices at above chance levels in pay-what-you-want situations (Lynn, Flynn and Helion, 2013), and are more likely to buy products priced with round numbers in field studies (e.g., Bray and Harris, 2006; Wieseke, Kolberg and Shons, 2016). Thus, it is possible that recommending round tip amounts will increase the attractiveness of the recommended tips and, therefore, the likelihood that people will select one of them and leave a tip.

In the closest test of this possibility of which we are aware, Edwards and List (2014) found no support for it; they found that suggesting a charitable donation of \$20 (vs. a donation of \$20.01, \$20.02, ..., \$20.08, or \$20.09) had no reliable effects on the proportion of subjects donating or on the average size of donations made in their study. However, their non-round suggestions were all larger than their round suggestion, so subjects (who tend to read left to right and to ignore the rightmost digits of prices) may have perceived the non-round suggestions of \$20.01 to \$20.09 no differently than the round suggestion of \$20 (see Thomas and Morwitz, 2005). Non-round suggestions that have a different left-most digit (e.g., \$2.99 vs. \$3.00) may be more perceptually distinctive from round suggestions and, therefore, may produce larger round vs. non-round suggestion effects on voluntary payments. Our data permitted us to assess this possibility and test the generalizability of Edwards and List's (2014) findings.

3. Setting, Study Design, and Data Description

Our data come from a laundry pick-up, cleaning, and delivery service (called "Washio") that operated in a total of six cities across several regions of the United States during the years 2014 to 2016. Communication between the company and its customers occurred via an app,

which customers downloaded and used to register for the service, place and pay for orders, receive information about pick-up and delivery times and drivers, receive invoices for completed orders, and rate completed transactions. On September 15, 2015, the firm added a function to its app that allowed consumers to electronically tip their delivery drivers using their accounts' credit card numbers. Prior to that date, any tipping that occurred was in cash. When tipping with the app, customers could choose one of three suggested tip amounts, could generate a custom tip amount, or could opt not to tip.

The company used each customer's randomly generated, permanent id number to assign him or her to one of the following eleven different sets of tip suggestions: a) \$2, \$4, \$6, b) \$3, \$4, \$5, c) \$3.95, \$4, \$4.05, d) \$3.99, \$5.99, \$7.99, e) \$4, \$6, \$8, f) \$5, \$8, \$10, g) 5%, 10%, 15%, h) 10%, 15%, 20%, i) 12%, 15%, 18%, j) 15%, 18%, 20%, or k) 15%, 20%, 25%. None of the tip suggestions were verbally labeled or described, so there were no differences within or across conditions in the explicit meanings of the different suggestion amounts (see Figure 1). The assignment to tip suggestion condition was permanent, so customers saw the same recommendations each time they patronized the service and were unlikely to be aware of their participation in an experiment. Customers were provided these tipping options when setting up or editing their account preferences (customers could save their tipping preferences and have them automatically applied to future transactions), placing or checking the status of an order (if there was no saved tip preference), and after receiving the electronic invoice for completed orders following delivery of the cleaned clothes (if no tip decision had been previously specified). Customers could also rate the service when they received their invoice (see Figure 1).

Since every customer was assigned to a (non-zero) tip suggestion condition, this natural field experiment does not allow us to study the impact of *introducing* tipping recommendations;

instead, we examine the effects of variation in the tip recommendations by comparing customers assigned to different groups.² The various tip recommendation conditions were selected by the firm for its own reasons, which were not disclosed to us and are not entirely obvious.

Nevertheless, they provided an opportunity to test the effects of several interesting variations in suggested tip amounts. In particular, we use these randomly assigned manipulations to examine the effects of providing: (i) large vs. small suggested tips (holding the asking range constant), (ii) dollar vs. percentage amounts of suggested tips (at various bill sizes), (iii) a broad vs. narrow range of suggested tips (holding the average suggested tip/contribution constant), and (iv) round vs. non-round (\$x.00 vs. \$x.99) suggested tips.

Washio provided data on all of its orders from customers not associated with corporate accounts over its entire period of operation – on 246,132 orders from 50,773 customers. However, 149,416 orders (from 39,504 customers) that were first invoiced before September 15, 2015 were dropped from the main analyses because those orders preceded the app's support of tipping and the customers' receipt of tip recommendations. These pre-15 September 2015 orders were used only to test the randomness of assignment to tip suggestion conditions as explained later. The post-15 September 2015 data included 2,145 orders from customers with corporate accounts. All of these observations were also dropped from analysis to ensure that customers were personally responsible for paying the bills and tips from all their orders. This left a total of 94,571 orders from 24,637 customers that were retained for the main analyses. However, the data provided by the company contained numerous missing values, so the number of observations

² Data on tipping prior to the app change is not available, so before-after comparisons of tipping are not possible. Furthermore, the company raised its prices substantially around the time it released the updated app supporting tipping and raised prices again a few months later, so the app update that supported tipping is confounded with price, which makes before-after comparisons of non-tipping variables difficult to interpret. However, the randomly assigned tip recommendations are not affected by this confound, so we focus on their effects in this paper.

varies across analyses depending on the variables involved.³ Descriptions of the variables examined in this study are presented in Table 1 and descriptive statistics for the key quantitative variables are presented in Table 2.

4. Analyses and Results

4.1. Analytic Approach

Our experiment had 11 conditions and 13 outcome measures, which created 715 possible paired-comparisons and a resulting problem with excessively high experiment-wide error rates. We addressed this problem in three ways. First, we began our analyses of each outcome with an omnibus test of the null hypothesis that the means for all the experimental conditions were equal to one another.⁴ Only if this omnibus test indicated that one or more of the means significantly differed from the others did we give much credence to subsequent significant paired-comparisons involving that outcome. If the omnibus test for an outcome was not significant, we still report the paired-comparisons, but regard those that were significant with some skepticism.

Second, we reduced the likelihood of making Type 1 errors by limiting our examination of paired-comparisons to only the nine comparisons that differed in a single meaningful way. Specifically, we focused on four paired-comparisons that differed only in recommended tip size (E vs A, H vs G, K vs H, and K vs G); two paired-comparisons that differed only in the range of recommended tip sizes (A vs B and H vs I); one pair-comparison that differed only in recommending even vs odd tip amounts – i.e., tip sizes with different left and right-most digits

³ Among the missing values in the data are 214 service ratings of zero that were recoded as missing because zero was outside the scale's range and it was unclear how those values were generated and saved or what they meant.

⁴ For the omnibus tests we regress the outcome variable on a full set of indicators for each tip suggestion condition and report the corresponding F (or χ^2) statistic for the model. This tested the null hypothesis that the mean for all of the conditions were equal to one another. In specifications containing multiple observations per customer, standard errors are adjusted for clustering at the customer-level. In specifications containing only one observation per customer (such as those for patronage frequency), our omnibus test statistic is equivalent to the model statistic reported from an analysis of variance.

that differed by only a penny (E vs D); and two paired-comparisons that differed only in expression of recommended tips as percentages vs dollars conditional on a bill size of \$40 (G vs A and H vs E).

Finally, we conducted a second set of paired comparisons that used List, Shaikh and Xu's (2016) method and code for adjusting analyses for multiple comparisons. This method adjusts for multiple comparisons involving both multiple treatments and multiple outcomes while taking into account inter-dependencies among the different treatments and among the different outcomes. The resulting tests control experiment-wide error rates while retaining more statistical power than traditional Bonferroni adjustments. However, this code does not accommodate multiple observations per customer or allow for clustering of error terms, so we performed these adjusted paired-comparisons using one randomly selected observation per customer. Only significant adjusted paired-comparisons involving an outcome measure with a significant omnibus test are regarded as compelling in our description and discussion of results below.⁵

4.2. Check on Randomization of Tip Suggestions

The eleven tip suggestion conditions were assigned to customers based on the customers' pseudo-random id numbers. The algorithm used did not ensure equal sample sizes across conditions, but its use of computer-generated pseudo-random ID numbers did give every customer the same chance as other customers to be in each of the conditions. Since all existing and new customers were assigned to tip suggestion conditions, we were able to compare pre-15 September 2015 orders across the conditions to which the customers placing those orders had been assigned with the expectation that no differences in customer satisfaction, patronage

⁵ Because of differing sample sizes across outcomes and to reduce computation time, we perform the tests for multiple comparisons using several sub-groupings of the dependent variables. The variables analyzed together are indicated in the notes to the corresponding tables. These analysis always account for the full set of 9 paired-comparisons that are the focus of our analysis.

frequency, and bill size across tip suggestion conditions should be evident before customers were exposed to those suggestions. As expected, omnibus tests using errors clustered within customer indicated that differences across assigned conditions in the mean pre-suggestion bill sizes, patronage frequency, and satisfaction ratings were all within chance levels (see Table 3).

4.3. Effects of Tip Suggestions on Tipping

Descriptive statistics for the measures of tipping behavior in this study are presented by tip suggestion condition in Table 4. Omnibus tests using errors clustered within customer, which are also reported in Table 4, indicated that the different tip suggestion conditions did result in reliable differences in the likelihood of tipping, size of non-zero tips, tip amount, and the variability in tip amounts (all p 's < .0001). Paired comparisons of selected conditions are also presented in Table 4 and are summarized in Table 8. We tested for mean differences in the outcome variables for sets of pairs for which the interpretations are most clear. For example, comparing condition A (\$2, \$4, \$6) to condition B (\$3, \$4, \$5) allows us to study the effect of the *range* of suggestions holding the average amount constant, while comparing condition A to condition E (\$4, \$6, \$8) allows us to study the effect of suggesting larger tip amounts while holding the range constant. The table indicates whether each pairwise difference in means was statistically significant at the 5% level when adjusting for multiple comparisons and multiple outcomes as well as using the naïve (unadjusted) test. In general, these comparisons indicated that both suggesting larger tip amounts (aka ask-size) and suggesting percentage vs. dollar tip amounts (aka percentage-ask) consistently increased the size of non-zero tips, average tip amounts (including tips of zero), and the variability in tip amounts. At the same time, a larger ask-size decreased the likelihood of leaving a tip, though the positive impact on non-zero tip size was large enough to outweigh the negative impact on tip likelihood, resulting in higher tip

amounts overall. We also found that suggesting a larger range of tip amounts and suggesting round vs. non-round tip amounts had no consistent or reliable effects on any of the tipping measures.

4.3.1. Additional analyses of ask-size effects

The effects of ask-size on tipping replicate similar effects reported by Haggag and Paci (2014). However, recommending larger tips reliably reduced tipping likelihood in this study only when the top recommendations were 15%, 20%, or 25% (condition K) and a post-hoc test comparing these significant ask-size effects with the others involving different conditions produced a χ^2 (1) of 9.62, $p < .002$. Thus, the negative ask-size effect on tipping likelihood appears to be limited to cases where one set of the recommended amounts exceeds familiar and normative levels.

To explore the robustness of the ask-size effect on tip amounts with as much statistical power as possible, a new variable called “ask-amount” was created to reflect the mean percentage tip requested in three otherwise comparable tip recommendation conditions – condition G (coded as 10), condition H (coded as 15) and condition K (coded as 20). Restricting the analysis to these three conditions, we regressed the tip amount on ask-amount and its interaction with several other variables; the results are presented in Table 5. As in the main analysis, a larger ask-amount produced larger tips. On average, an increase in the mean tip suggestion of one percentage point led to an \$0.17 increase in tips. This effect was moderated by some but not all of the variables we examined. It was reliably stronger the larger the bill size, which reinforces the idea that you get what you ask for because the dollar size of a given percentage increases with bill size.

The effect of ask-amount on tip amount was substantially larger for those with a saved tip preference, though it remained positive and statistically significant even for those without a saved tip preference. Unfortunately, it is difficult to precisely interpret these effects because tip recommendations also affected the tendency to save a tip preference (see Table 7) and therefore the ask-amount-by-saved-tip-preference interaction is likely to be endogenous. This interaction may result from a desire to minimize the cognitive effort of tipping, which might incline people both to go along with tip recommendations and to save their tip preference.⁶

We found no significant interaction between ask-amount and the median income of the customer's zip code, or between ask-amount and customer satisfaction with the service. We also investigated whether the impact of the tip suggestion varied with the number of times the customer was exposed to the treatment. To do so we interacted ask-amount with post-treatment order number, where, for example, order number 1 is the first order the customer placed after 15 September 2015. In order to control for compositional effects, we restrict the sample to only those customers who placed at least four orders and we look at the effects across only those first four orders.⁷ There was a positive and statistically significant interaction between ask amount and order number over the first four orders, which provides suggestive evidence that the magnitude of the treatment effect may have increased over time (column 6 of Table 5). However, one limitation of this analysis is that an app update was required in order to enable the tipping functionality and tip suggestions for customers using mobile devices. Unfortunately we do not know precisely when each customer updated the app, so these results are confounded with the

⁶ Consistent with this possibility, the tendency to use a default tip and to save a tip preference were positively related ($B = 1.78$, $S.E. = .03$, $p < .001$) in a binomial logistic regression of default tip on saved preference using robust standard errors clustered within customer.

⁷ Because not all customers place the same number of orders, the number of customers placing order number x drops as x increases. Our sample restriction ensures that the pool of customers does not change as order number changes.

fact that customers were less likely to have been exposed to the treatment during earlier orders.⁸ To attempt to disentangle these effects, we ran two additional specifications. First, we repeated the specification from Table 5 using the size of non-zero tips as the outcome variable, which ensures that the sample is restricted to customers who have been exposed to the tip suggestions; and second, we restricted the analysis to customers who placed their first post-treatment order after 01 November 2015, increasing the likelihood that these customers will have the tipping option available at their first order. In both cases, we again found positive and significant effects, suggesting that the observed differences by order number may not be fully explained by delays in updating the app.⁹

4.3.2. Additional analyses of percentage-ask effects

To explore the processes underlying the percentage-ask effect, two new indicator variables – G%vsA\$ and H%vsE\$ – were created to contrast the two types of requests holding other ask characteristics constant. G%vsA\$ was coded as 0 for condition A (\$2, \$4, \$6) and 1 for condition G (5%, 10%, 15%). Similarly, H%vsE\$ was coded as 0 for condition E (\$4, \$6, \$8) and 1 for condition H (10%, 15%, 20%). For both pairs of conditions, the dollar and percentage recommendations amounted to the same ask-size when bill size was exactly \$40, while the dollar recommendations amounted to a larger ask-size below \$40 and the percentage recommendations amounted to a larger ask-size above \$40.

⁸ For customers who did not update the app, tip amounts are recorded as zero rather than missing. This limitation presents a problem for our analysis of any dynamic effects. However, it should not affect the validity of our other estimates as there is no reason to believe that the propensity to update the app varied across the (randomly assigned) treatment groups.

⁹ For the first specification, the coefficient on the interaction term was 0.0564 with a standard error of 0.0199; and for the second, the coefficient was 0.0586 with a standard error of 0.0231, which is similar in magnitude to the corresponding coefficient in Table 5 using the larger sample.

Consistent with the main analyses, both percentage-ask indicators had positive effects on tip amount (see Table 6). Since roughly 80 percent of the observations involved bill amounts over \$40, these percentage-ask effects could be due to differences in the recommended tip amounts. In other words, they could be disguised ask-size effects. Consistent with this possibility, the effects of both variables grew reliably stronger as bill size, and the resulting difference in the ask-size of the percentage vs. dollar recommendations, increased. However, the effects of G%vsA\$ and H%vsE\$ were not reliably reversed at bill sizes below \$40 as would be expected if percentage vs. dollar asks had only ask-size effects.

To identify potential percentage vs. dollar ask effects that were independent of ask-size, we examined the effects of G%vsA\$ and H%vsE\$ in a sub-sample with bill sizes between \$39.50 and \$40.50, where the difference in ask-size across percentage vs. dollar asks was minimal (see Table 6). H%vsE\$ had no reliable effects on tip amount in this sub-sample analysis, but G%vsA\$ had a reliably positive effect (see Table 6). Perhaps subjects' familiarity with tipping 15 to 20% in restaurants made the 5% and 10% options in condition G seem too small and encouraged selection of the 15% option, while subjects' familiarity with all the options of \$2, \$4 and \$6 in condition A created no comparable encouragement to select the \$6 option. This would explain why H%vsE\$ had no reliable effects because it involved fewer sub-normative tip percentage options.

4.4 Effects of Tip Suggestions on Customer Satisfaction, Patronage Frequency and Spending

Descriptive statistics for the measures of customer satisfaction, patronage frequency and spending in this study are presented by tip suggestion condition in Table 7. Omnibus tests using errors clustered within customer, which are also reported in Table 7, indicated that the different tip suggestion conditions did not reliably affect average customer satisfaction, patronage

frequency or spending. Nor did the different tip suggestion conditions affect variability in customer spending. They did affect variability in customer satisfaction and patronage frequency – variability in customer satisfaction was slightly lower for tip condition G than for the other conditions and variability in patronage frequency was slightly lower in conditions E and J than the other conditions. However, those effects were idiosyncratic and neither theoretically or economically important.

Simple paired comparisons did find that satisfaction ratings were higher in condition G (5%, 10%, 15%) than in conditions H (10%, 15%, 20%) and K (15%, 20%, 25%), but the difference in satisfaction ratings was small in magnitude (less than .1 on a 5 point scale) and did not remain significant in analyses adjusting for multiple comparisons. Furthermore, satisfaction ratings did not differ between conditions H and K or between conditions A and E. Thus, the paired comparisons of tip recommendations that affected tipping did not reliably affect central tendencies or variabilities in customer satisfaction, patronage frequency or spending (see Tables 7 and 8). This means that increasing the amounts customers tipped did not affect these outcomes either, which suggests that none of the self-perception/justification, warm-glow, or value-perception processes observed in other contexts operated here.

We next examine whether our failure to find evidence that tip recommendations affected the tipper's satisfaction could be attributable to problems with our measurement of satisfaction. First, customers did not have to supply satisfaction ratings and it is possible that only those who were satisfied chose to provide the ratings. This self-selection process could have hidden any negative effects of tip recommendations on satisfaction ratings. However, tip recommendations did not affect the likelihood of providing satisfaction ratings (see Table 7), so if completion of

the ratings is itself a sign of satisfaction as this argument assumes, then the null results involving mean satisfaction ratings were replicated using this new measure.

Second, we consider the possibility that satisfaction ratings were so high that the resulting restriction of range may have hidden the true effects of tip recommendations on those ratings. However, those satisfaction ratings were not too high or restricted to prevent them from being reliably related to tip amount ($B = .28$, $S.E. = .03$, $p < .001$), patronage frequency ($B = .44$, $S.E. = .05$, $p < .001$), or bill size ($B = -1.40$, $S.E. = .23$, $p < .001$) in separate regressions of satisfaction on these variables using robust standard errors clustered with customer. Thus, if restriction of range in the satisfaction ratings hid the true effects of tip recommendations on satisfaction, those true effects must have been very small. Also arguing for at best small and inconsequential true effects of tip recommendations on satisfaction is the fact that those recommendations did not affect re-patronage frequency or spending either.¹⁰

4.5 Effects of Tip Suggestions on Consumer Use of Default Tip Amounts

The proportion of orders on which customers tipped one of the recommended or default options is presented by tip recommendation condition in the final column of Table 4. Omnibus tests using errors clustered within customer, which are also reported in Table 4, indicated that the different tip suggestion conditions did affect use of the suggested tip amounts. Paired comparisons of selected conditions are also presented in Table 4 and are summarized in Table 8.

¹⁰ Given the skewness of our satisfaction variable—the highest rating is given for about 74% of orders—we also confirm that the results are robust to alternative specifications. Instead of OLS, we used an ordered probit model to estimate the coefficients and standard errors for the omnibus test and naïve paired comparisons, and we find similar results. In addition, rather than using the 5-point scale, we constructed an indicator for whether the rating equals 5, the highest value. Again the results were similar using this measure: the omnibus test was not significant; condition G again stood out as having slightly higher satisfaction, but those results were not significant after controlling for multiple comparisons.

As might be expected, these tests indicated that consumers were more likely to use one of the recommended or default tip options when the suggested tips were smaller and had a larger range.

5. Discussion

5.1. Summary of Key Findings

Among the noteworthy findings from our analysis of this natural field experiment on the effects of various tip recommendations are that: (i) arithmetically equivalent percentage vs. dollar tip recommendations increased tip amounts in the case where some of the percentages recommended were sub-normative, (ii) increasing the range of recommended tips increased selection/use of one of the recommended amounts, but did not affect the amount tipped, and (iii) round and non-round tip recommendations (with different left-most as well as right-most digits) did not produce different levels of tipping. However, the most interesting and important findings are that recommending larger tip amounts decreased the likelihood of tipping (when the recommendations exceeded familiar and normative amounts) and increased the size of tips that were left as well as overall tip revenues, but had no comparable or opposite effects on customer satisfaction, re-patronage frequency, or spending.¹¹ These latter results have implications for (i) our confidence in the existence, or not, of the effects, (ii) possible boundary conditions for the processes hypothesized to underlie the effects, and (iii) the advisability of using larger asks to boost voluntary payments. These issues are discussed below.

¹¹ These effects on tipping reflect only app-based tipping. Some customers may have given drivers cash tips rather than tipping through the app and we have no way of seeing how the tip recommendations affected these cash tips. However, anecdotal evidence suggests that cash tips were relatively rare. In addition, any cash tips were probably given instead (not in addition to) app-based tips. Thus while the magnitude of our estimated negative effect on tip likelihood may be overstated in the presence of cash tipping, our estimated impacts on tip amount are likely to be a lower bound for the true effect. Most importantly, the possibility that tip recommendations could have affected cash tips as well as app-based tips does not challenge our finding that tip-enhancing recommendations do not affect customer satisfaction, re-patronage frequency and spending.

5.2. Confidence in Ask-Size Effects

It has been noted that replication rates in the psychological sciences are disappointingly low (Open Science Collaboration, 2015) and that published effects are more likely to be false than true (Ioannidis, 2005), so readers may wonder about the confidence they should place in our conclusions. Fortunately, the large sample size in this study – involving 94,571 orders from 24,637 customers – provides high levels of power and, therefore, tight confidence intervals. For example, the 95% confidence intervals for our 9 unadjusted pair-wise comparisons of mean tip amounts ranged from $\pm \$0.18$ to $\pm \$0.54$ (on a measure with an overall standard deviation of \$4.85) and those for our 9 unadjusted pair-wise comparisons of mean satisfaction ratings ranged from ± 0.05 to ± 0.07 (on a 5-point scale with an overall standard deviation of .97).

For a more complete perspective on the level of confidence that our findings warrant, we use a formula described by Maniadas, Tufano and List (2014) to calculate the post-study probability (PSP) that our positive conclusions about ask-size effects are true given various priors regarding those probabilities. For positive effects, Maniadas et. al. argue that “PSP is equal to the number of true associations which are declared true divided by the number of all associations which are declared true: $PSP = (1 - \beta)\pi / ((1 - \beta)\pi + \alpha(1 - \pi))$ ”, where π is the fraction of associations that are true (or are assumed to be true a-priori), $1 - \beta$ is the statistical power of the study, and α is the alpha level used in the study. We calculated the post-hoc powers of the pair-wise comparisons of ask-size in our study using the standard formula for the power of an inference about the means from two independent samples. In these calculations, we used the number of customers in each condition rather than the number of observations, as the sample

sizes, so our power estimates are likely to be conservative.¹² Nevertheless, all of the comparisons had a post-hoc power of 0.99, so that value was used to calculate the post study probabilities of ask-size effects on tip amounts of the magnitudes we observed given various priors (see Table 9). The post study confidence that should be placed on these effects depends on the level of confidence in the effects held prior to the study, but is substantially greater than those priors. For a prior of 0.01, the post study probabilities of the effects increase 1,600%, and for a prior of 0.5, they increase 90%. The priors would have to be lower than 0.05 to bring the post study probabilities below 0.5.

Maniadas, Tufano and List (2014) do not provide a formula to calculate the post-study probability of a null result (PSPN). Nor could we find one that suited our needs elsewhere. However, using the same logic as previously, we argue that PSPN is equal to the number of true null-effects which are declared null divided by the number of all effects which are declared null: $PSPN = (\pi'(1 - \alpha)) / ((\pi'(1 - \alpha) + \beta(1 - \pi')))$, where true null-effects are defined as absolute differences between condition means of less than the smallest amount considered meaningful, π' is prior probability of a true null-effect, β is the chance of getting a non-significant result if the smallest effect considered meaningful is true, and α is alpha level. We used this formula along with our study sample sizes (numbers of customers per condition) and standard deviations, to calculate the post study probabilities that our ask-size null-effects on satisfaction ratings were true, where true null-effects are defined as those whose absolute value is less than 0.1 (see Table 9).¹³ Again, the post study confidence that should be placed on these null-effects depends on the

¹² No software or formula known to us permits a better or more appropriate assessment of power because our data included clusters of unequal sizes.

¹³ The effects of HvsG and KvsG were not significant in pair-wise comparisons adjusting for multiple comparisons, but did have reliable (though trivial) effects on satisfaction ratings in unadjusted pair-wise comparisons. Thus, our classification of these effects as null is subject to challenge. This is a problem inherent to the application of binomial

level of confidence in the null-effects held prior to the study, but is substantially greater than those priors. For a prior of 0.01, the post study probabilities of these null-effects increase 300 to 400%, and for a prior of 0.5, they increase 46 to 56%. Priors as low as 0.25 result in post study probabilities of at least some of these null-effects to exceed 0.5.

5.3. Potential Boundary Conditions for Hypothesized Underlying Processes

The finding that larger tip recommendations affected customers' tipping behavior but not their satisfaction, re-patronage frequency or spending suggests that neither the processes underlying the tip recommendation effects on tipping nor consumer tipping itself affect these other customer outcomes. In particular, the finding that asking for super-normative tips decreased tipping likelihood but not customer satisfaction, re-patronage and spending (or increase the variability in these measures) suggests that the former effect is not due to strong anger or other negative affect that might be expected to also impact these latter outcome variables. Previous research has found that unsolicited recommendations can lead to negative affect and dissatisfaction as well as reactive or contrary responses (Fitzsimons and Lehmann, 2004), but that does not seem to have happened in this case. Why our super-normative tip recommendations appear to have created reactance (i.e., negative effects on compliance) without negative affect impacting satisfaction and other outcomes is not clear. Perhaps the fact that the tip recommendations benefited drivers rather than the firm softened consumers' emotional reactions to the firm, but not their behavioral reactions to the recommendations themselves. Developing and testing more ideas about the conditions under which recommendations produce

logic to continuous data, but does not detract from the value of the approach in cases with p-values farther from the alpha level.

reactance with and without negative affect toward the recommender is an interesting issue left to future research.

The fact that inducing people to tip more did not affect the tippers' satisfaction ratings, re-patronage frequency or subsequent spending also suggests the self-perception/justification, warm-glow, and value-perception processes observed in other contexts did not operate here.¹⁴ Characteristics of the current context that might be boundary conditions responsible for the failure to find evidence of these processes are discussed below.

5.3.1. Self-perception/justification and warm-glow effects of tipping

Voluntary compliance with small requests has been shown to evoke self-perception and consistency processes that increase subsequent compliance with larger requests (Burger, 1999) and voluntary pro-social behavior has been shown to generate a positive feeling or “warm glow” that enhances customer satisfaction and re-patronage intentions (Giebelhausen, Lawrence, Chun and Hsu, 2017). Given this research and the fact that tipping is a voluntary behavior that benefits another person, it seemed likely that tipping would evoke self-perception/justification and warm-glow processes that increase customer satisfaction, re-patronage and spending, but such effects

¹⁴ Of course, it is possible that the true effects of tip recommendations on satisfaction, re-patronage and spending were hidden in our study by other, offsetting processes. For example, some tip recommendations decreased tipping likelihood while increasing the size of those tips left and it is possible that the downstream consequences of these opposite effects on tipping canceled one another out. However, those pair-wise comparisons that produced opposite effects on tipping likelihood and tip size should have increased the variability in customer satisfaction, re-patronage, and spending if tipping behavior affects these later outcomes, and we found no such effects. Furthermore, a number of the pair-wise comparisons produced effects on tip size only (with no effects on tipping likelihood) and yet these comparisons did not produce reliable effects on satisfaction, re-patronage or spending either. Alternatively, tipping may have both positive effects on satisfaction, re-patronage and spending through self-perception/justification and warm-glow processes and negative effects through value-perception processes, with these opposing processes and effects offsetting one another. However, this possibility requires a precise balance of tipping-induced warm-glow, self-perception/justification, and value-perception effects across many different tip recommendation manipulations and outcome measures that is highly unlikely to occur by chance. Moreover, even in the unlikely event that such a robust balance among these processes and effects did exist, we would still conclude that tipping does not affect these outcomes and our practical implications would remain the same.

were not observed. Something about the current study conditions appears to interfere with these processes and their effects.

Self-perception/justification and warm-glow processes depend on the actors' attributions for their voluntary behavior. For example, both self-perception and warm-glow effects are diminished when external incentives for the voluntary behaviors diminish the behaviors' signaling about the self (Burger and Caldwell, 2003; Dillard, Hunter and Burgoon, 1984; Giebelhausen and Chun, 2017; Giebelhausen, Chun, Cronin and Hult, 2016). Thus, it is possible that these processes and their effects are limited to contexts where there is a dominant internal attribution for behavior. The current study context may not have met this condition, because tipping is driven by many motives, such as desires to reward good service, help servers, buy future service, buy social esteem, and comply with social norms (Lynn, 2015a, 2015b, 2016). The multi-faceted nature of motivations for tipping may interfere with the self-attributions underlying self-perception/justification and warm-glow processes and, thereby, diminish the effects of tipping on subsequent attitudes and behaviors. Since many other real-world behaviors are similarly driven by multiple motivations, such a boundary condition for self-perception/justification and warm-glow processes and their effects is potentially important and deserves further investigation.

5.3.2. Value-perception effects of tipping

Our finding that increasing customers' tip expenditures did not adversely affect their satisfaction, re-patronage frequency or spending suggests that those customers were not very sensitive to the costs associated with tipping. This conclusion is further supported by the fact we found no marginal decrease in the positive relationship between tip amounts and bill size as bill size increased; that is, that the relationship between tip amounts and bill size was relatively

linear, as shown in Figure 2. When we trim the 5 observations where bill size was above \$750 (to reduce the influence of outliers, a regression of tip amount on bill size and bill size squared results in a quadratic term that is small and statistically insignificant ($B_{\text{quadratic}} = -.00001$, S.E. clustered within customer = .00002, $p > .50$). In other words, customers tipped roughly the same proportion of their bill sizes as those bill sizes increased even though the costs of doing so increased with bill size. Thus, customers appear relatively insensitive to a broad range of costs associated with tipping.

The insensitivity to tipping costs in this study is interesting because it suggests that consumers may be less sensitive to voluntary price increases than to involuntary ones. There are three potential reasons for such a difference in price sensitivity. First, voluntary price increases are more easily avoided than are involuntary price increases, because voluntary pricing allows price-sensitive consumers to simply choose a lower payment while involuntary pricing requires them to either pay the higher amount or forgo the intended purchase. Second, paying larger voluntary prices reflects or signals the givers' wealth and generosity more than does paying larger involuntary prices and this reputational benefit may lessen the pain of paying. Finally, voluntary pricing empowers consumers both objectively and subjectively (Barone, Bae, Qian, and d'Mello, 2017) and feelings of power may decrease price sensitivity by focusing consumers' attention on the acquisition of rewards and away from the prevention of pain and loss (Keltner, Gruenfeld and Anderson, 2003; Yang, Li, Guo, Fan and He, 2017). Of course, the current data only support insensitivity to voluntary price increases, not comparatively lower sensitivity to voluntary than to involuntary price increases. However, the former makes the latter more likely. Furthermore, other research finding that the perceived expensiveness of restaurants is affected more by menu prices than by expected tip amounts (Lynn and Wang, 2013; Lynn, 2017)

provides additional support for the idea that consumers may be less sensitive to voluntary price increases than to involuntary ones.

5.4. Practical Implications

Our findings also have practical implications for those individuals and organizations seeking to increase tips or other voluntary contributions. In particular, our findings suggest that increasing consumer tipping (by itself) is neither a viable means of increasing customer satisfaction, re-patronage and spending nor a threat to these outcomes. As long as the means used to alter tipping do not directly affect customer satisfaction, then tips can be increased or decreased with few other consumer consequences. Furthermore, our results indicate that recommending larger tip amounts, at least within the \$2 to \$10 or 5% to 25% ranges studied here, is a safe means of increasing the amounts customers leave. Recommending larger tips increases tip revenues without long term adverse effects on attitudes or behavior even when those recommendations cause some recipients to not give in the short term.

The app-based laundry pick-up, cleaning, and delivery service we studied shares many characteristics with other app-based delivery services such as GrubHub, Instacart, Postmates, and Amazon Flex but is quite different from more traditional tipping contexts. For example, the current study context involves less server-customer contact and less service customization than does tipping in restaurants, bars, hair salons and hotels. However, the greater server-customer contact and service customization in more traditional tipping contexts seems likely to decrease rather than increase customer irritation and anger with larger tip recommendations, so our conclusion that it is safe to increase the size of recommended tips should still apply to those traditional tipping contexts.

Our conclusion that increasing the tips customers leave will not lower customer's perceptions of value, satisfaction, re-patronage frequency and subsequent spending – in contrast to expectations based on the law of demand – should also apply across tipping contexts, because the costs of tipping more are the same across those contexts. Arguably, the customers in our study could be wealthier than most consumers and this could limit the generalizability of our conclusions about sensitivity to the costs of tipping. However, other research finding that there is no quadratic trend in the positive relationship between restaurant tip amounts and bill size (Lynn and Sturman, 2003) as well as the previously mentioned research finding that perceptions of restaurants' expensiveness are affected more by menu prices than by expected tip amounts (Lynn and Wang, 2013; Lynn, 2017) suggest that insensitivity to the costs of tipping is not limited to our sample and study context.

Things are a little different with respect to our conclusion that tipping does not enhance customer satisfaction and re-patronage. Reasonable arguments can be made that the unusual features of the current study context may have diminished self-perception/justification and warm-glow processes. For example, less server-customer contact may reduce altruistic motives for tipping and, hence, warm-glow processes, while less service complexity/customization may reduce reward motives for tipping and, hence, self-perception/justification processes. Thus, it is possible that our failure to find tip recommendation effects on customer attitudes and patronage speak only to the effects of tipping for app-based delivery services and that such effects would be stronger in more traditional tipping contexts. This possibility is worth testing in future research. However, if a multiplicity of motivations for tipping does undermine self-perception/justification and warm-glow processes as discussed previously, then even research conducted in more traditional tipping contexts is unlikely to find strong tipping effects on customer attitudes and

behavior because tipping is driven by many motives in most service contexts (Lynn 2015b, 2016). Positive but weak correlations of restaurant tip amounts with customer service ratings and patronage frequency in the existing tipping literature (see Lynn and McCall, 2000) lend credence to the generalizability of our null tipping effects on satisfaction and re-patronage, because those weak correlations undoubtedly reflect satisfaction and patronage frequency effects on tipping and provide little room for the reverse effects as well.

A final caution regarding the generalizability of our findings and their practical implications concerns labor market dynamics. As previously mentioned, our results suggest that if tips are paid out directly to workers, it is possible for firms to increase the earnings of tipped workers with no loss in revenue or profits. Firms may also want to capture some of this revenue for themselves, for example by reducing their employee compensation in response to the worker's tip earnings. In fact this practice has been the subject of some controversy in settings similar to the one that we study (see Houk, 2019). Obviously, the current data do not speak to all the issues affecting the advisability of this tip-skimming practice. Importantly, they do not even speak to the more limited issue concerning the advisability of asking for larger tips when those larger tips are known or thought to be skimmed by firms in the form of lower wages. Our results were observed in a context in which service workers were presumed to benefit from the tips given. It remains an open question how applicable the results from this paper would be in a situation in which firms are perceived as the primary beneficiaries of larger tips, because they are believed to lower employee wages as employee tip income increases.

5.5. *Conclusions*

The main finding from this paper is that larger suggested tip sizes increased the amount of tips received while having no impact on overall customer satisfaction, re-patronage, or spending. Because the tip suggestions were randomized across customers who were unaware of their participation in a study and because the sample size was large, we can have high confidence in the internal validity and the statistical-inference validity of the findings. Of course, the specific nature of the study context – an internet app-based laundry delivery service – leaves open a question about the findings' generalizability to other contexts that should be examined in future research. Nevertheless, as discussed above, these findings provide insights to scholars interested in tipping as well as to those interested in more general value-perception, self-perception, self-justification, and warm-glow processes. They also provide insights to firms thinking about managing their employees' incomes and their customers' tipping behavior via tip recommendations and other tipping policies.

References

- Al-Ubaydli, O., and List, J. A. (2013). *On the generalizability of experimental results in economics: With a response to Camerer* (No. w19666). National Bureau of Economic Research.
- Aronson, E. (1969). The theory of cognitive dissonance: A current perspective. *Advances in experimental social psychology*, 4, 1-34.
- Azar, O. H. (2007). The Social Norm of Tipping: A Review 1. *Journal of Applied Social Psychology*, 37(2), 380-402.
- Azar, O.H. (2010). Tipping motivations and behavior in the U.S. and Israel. *Journal of Applied Social Psychology*, 40 (2), 421-457.
- Azar, O. (2011). Business strategy and the social norm of tipping. *Journal of Economic Psychology*, 32, 515-525.
- Barone, M. J., Bae, T. J., Qian, S., & d'Mello, J. (2017). Power and the appeal of the deal: how consumers value the control provided by Pay What You Want (PWYW) pricing. *Marketing Letters*, 28(3), 437-447.
- Becker, C., Bradley, G. T., & Zantow, K. (2012). The underlying dimensions of tipping behavior: An exploration, confirmation, and predictive model. *International Journal of Hospitality Management*, 31(1), 247-256.
- Bray, J.P., and Harris, C. (2006). The effect of 9-ending prices on retail sales: A quantitative UK based filed study. *Journal of Marketing Management*, 22, 601–617.
- Burger, J. M. (1999). The foot-in-the-door compliance procedure: A multiple-process analysis and review. *Personality and Social Psychology Review*, 3(4), 303-325.
- Burger, J. M., & Caldwell, D. F. (2003). The effects of monetary incentives and labeling on the foot-in-the-door effect: Evidence for a self-perception process. *Basic and applied social psychology*, 25(3), 235-241.

- Carr, C. L. (2007). The FAIRSERV model: Consumer reactions to services based on a multidimensional evaluation of service fairness. *Decision Sciences*, 38(1), 107-130.
- Cialdini, R. B., & Schroeder, D. A. (1976). Increasing compliance by legitimizing paltry contributions: When even a penny helps. *Journal of Personality and Social Psychology*, 34(4), 599.
- Clee, M. A., and Wicklund, R. A. (1980). Consumer behavior and psychological reactance. *Journal of Consumer Research*, 6(4), 389-405.
- De Bruyn, A., & Prokopec, S. (2013). Opening a donor's wallet: The influence of appeal scales on likelihood and magnitude of donation. *Journal of Consumer Psychology*, 23(4), 496-502.
- Desmet, P. (1999). Asking for less to obtain more. *Journal of Interactive Marketing*, 13(3), 55-65.
- Dillard, J.P., Hunter, J.E. and Burgoon, M. (1984). Sequential-request persuasive strategies: Meta-analysis of foot-in-the-door and doo-in-the-face. *Human Communication Research*, 10 (4), 461-488.
- Doob, A. N., & McLaughlin, D. S. (1989). Ask and You Shall be Given: Request Size and Donations to a Good Cause 1. *Journal of Applied Social Psychology*, 19(12), 1049-1056.
- Dubois, D., Rucker, D. D., & Galinsky, A. D. (2010). The accentuation bias: Money literally looms larger (and sometimes smaller) to the powerless. *Social Psychological and Personality Science*, 1(3), 199-205.
- Edwards, J. T., & List, J. A. (2014). Toward an understanding of why suggestions work in charitable fundraising: Theory and evidence from a natural field experiment. *Journal of Public Economics*, 114, 1-13.
- Estelami, H.(1999).The computational effect of price endings in multi-dimensional price advertising. *Journal of Product and Brand Management*,8, 244–256.
- Fazio, R. H. (1987). Self-perception theory: A current perspective. In *Social influence: the Ontario symposium* (Vol. 5, pp. 129-150). Lawrence Erlbaum Associates, Inc Hillsdale, NJ.
- Fitzsimons, G.J. and Lehmann, D.R. (2004). Reactance to recommendations: When unsolicited advice yields contrary responses. *Marketing Science*, 23 (1), 82-94.

- Gawronski, B. (2012). Back to the future of dissonance theory: Cognitive consistency as a core motive. *Social Cognition*, 30(6), 652.
- Giebelhausen, M., Chun, H., Cronin, J.J. and Hult, T. (2016). Adjusting the Warm Glow Thermostat: How Incentivizing Participation in Voluntary Green Programs Moderates Their Impact on Service Satisfaction, *Journal of Marketing*, 80(4), 56-71.
- Giebelhausen, M. and Chun, H. (2017). Replicating and Extending Our Understanding of How Hospitality Managers Can Adjust the “Warm Glow Thermostat, *Cornell Hospitality Quarterly*, 58(2), 122- 133.
- Giebelhausen, M., Lawrence, B., Chun, H. and Hsu, L. (2017), The Warm Glow of Restaurant Checkout Charity, *Cornell Hospitality Quarterly*, 58 (4), 329-341.
- Guido, G., and Peluso, A. (2004). Consumers’ perception of odd-ending prices with the introduction of the Euro. *Journal of Product and Brand Management*, 13, 200–210.
- Haggag, K. and Paci, G. (2014). Default tips. *American Economic Journal: Applied Economics*, 6 (3), 1-19.
- Houck, B. (2019). Instacart is revising its outrageous tipping policy, following public outcry. <https://www.eater.com/2019/2/6/18214354/instacart-delivery-worker-pay-tipping-policy-doorDash> , accessed 3/19/19.
- Ioannidis, J. (2005). Why most published research findings are false. *PLoS Medicine*, 2 (8), 1418-1422.
- Keltner, D., Gruenfeld, D. H., & Anderson, C. (2003). Power, approach, and inhibition. *Psychological Review*, 110(2), 265.
- List, J. A., Shaikh, A. M., & Xu, Y. (2016). Multiple hypothesis testing in experimental economics. *Experimental Economics*, 1-21.
- Lynn, W. M. (2006). Tipping in restaurants and around the globe: An interdisciplinary review. Chapter 31, pp. 626-643. IN Morris Altman (Ed.) *Handbook of Contemporary Behavioral Economics: Foundations and Developments*, M.E. Sharpe Publishers.

- Lynn, M. (2009). Individual differences in self-attributed motives for tipping: Antecedents, consequences and implications. *International Journal of Hospitality Management*, 28(3), 432-438.
- Lynn, M. (2015a). Service gratuities and tipping: A motivational framework. *Journal of Economic Psychology*, 46, 74-88.
- Lynn, M. (2015b). Explanations of service gratuities and tipping: Evidence from individual differences in tipping motivations and tendencies. *Journal of Behavioral and Experimental Economics*, 55, 65-71.
- Lynn, M. (2016). Motivations for tipping: How they differ across more and less frequently tipped occupations. *Journal of Behavioral and Experimental Economics*, 65, 38-48.
- Lynn, M. (2017). Should U.S. restaurants abandon tipping? A review of the issues and evidence. *Psychosociological Issues in Human Resource Management*, 5(1), 120-159.
- Lynn, M. (2018). Are published techniques for increasing service-gratuities/tips effective? P-curve and R-indexing the evidence. *International Journal of Hospitality Management*, 69, 65-74.
- Lynn, M., Flynn, S. M., and Helion, C. (2013). Do consumers prefer round prices? Evidence from pay-what-you-want decisions and self-pumped gasoline purchases. *Journal of Economic Psychology*, 36, 96-102.
- Lynn, M. and McCall, M. (2000). Gratitude and Gratuity: A Meta-Analysis of Research on the Service-Tipping Relationship. *Journal of Socio-Economics*, 29, 203-214.
- Lynn, M., & Sturman, M. C. (2003). It's simpler than it seems: an alternative explanation for the magnitude effect in tipping. *International Journal of Hospitality Management*, 22(1), 103-110.
- Lynn, M. and Wang, S. (2013). The indirect effects of tipping policies on patronage intentions through perceived expensiveness, fairness and quality. *Journal of Economic Psychology*, 39, 62-71.
- Lynn, M. and Withiam, G. (2008). Tipping and Its Alternatives: Business Considerations and Directions for Research. *Journal of Services Marketing*, 22 (4), 328-336.
- Maniadis, Z., Tufano, F. and List, J.A. (2014). One swallow doesn't make a summer: New evidence on anchoring effects. *American Economic Review*, 104 (1), 277-290.

- Namasivayam, K., & Upneja, A. (2007). Employee preferences for tipping systems. *Journal of Foodservice Business Research*, 10(2), 93-107.
- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349 (6251), aac4716, -1
- Robins, R. W., and John, O. P. (1997). The quest for self-insight: Theory and research on accuracy and bias in self-perception. Hogan, Robert (Ed); Johnson, John A. (Ed); Briggs,
- Schibrowsky, J. A., & Peltier, J. W. (1995). Decision frames and direct marketing offers: A field study in a fundraising context. *Journal of Direct Marketing*, 9(1), 8-16.
- Schwartz, Z. (1997). The Economics of Tipping: Tips, Profits and the Market's Demand—Supply Equilibrium. *Tourism Economics*, 3(3), 265-279.
- Seiter, J., Brownlee, G. and Sanders, M. (2011). Persuasion by way of example: Does including gratuity guidelines on customers' checks affect restaurant tipping behavior? *Journal of Applied Social Psychology*, 41 (1), 150-159.
- Shamir, B. (1983). A note on tipping and employee perceptions and attitudes. *Journal of Occupational Psychology*, 56(3), 255-259.
- Shy, O. (2015). Do tips increase workers' income? *Management Science*, 61 (9), 20141-2051.
- Stephen R. (Ed). (1997). *Handbook of personality psychology* , (pp. 649-679). San Diego, CA, US: Academic Press.
- Thomas, M., & Morwitz, V. (2005). Penny wise and pound foolish: the left-digit effect in price cognition. *Journal of Consumer Research*, 32(1), 54-64.
- Varki, S., and Colgate, M. (2001). The role of price perceptions in an integrated model of behavioral intentions. *Journal of Service Research*, 3(3), 232-240.
- Weyant, J. M., & Smith, S. L. (1987). Getting More by Asking for Less: The Effects of Request Size on Donations of Charity 1. *Journal of Applied Social Psychology*, 17(4), 392-400.
- Wieseke, J., Kolberg, A., and Schons, L. M. (2016). Life could be so easy: the convenience effect of round price endings. *Journal of the Academy of Marketing Science*, 44(4), 474-494.

Yang, W., Li, Q. Guo, M. Fan, Q. and He, Y. (2017). The effects of power on human behavior: The perspective of regulatory focus. *Acta Psychologica Sinica*, 49 (3), 404-415.

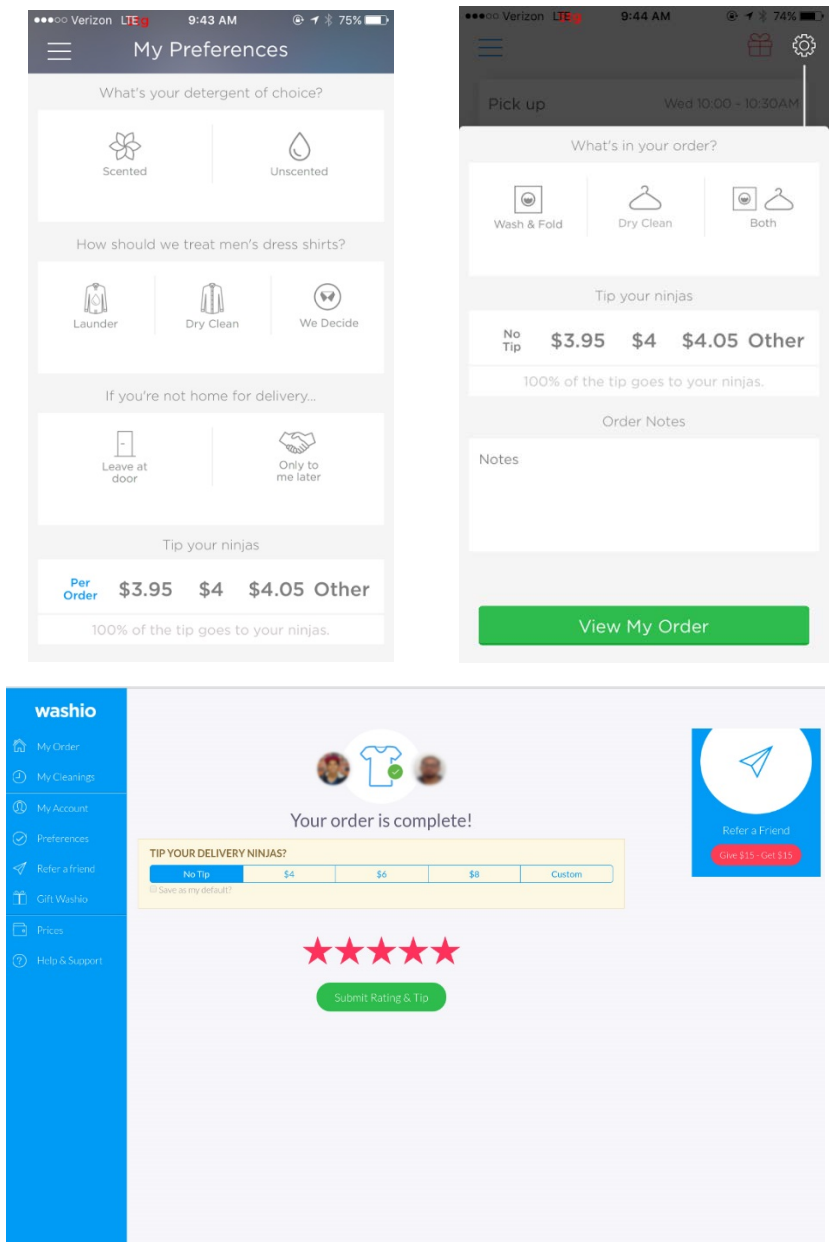


Figure 1. Examples of how the tip recommendations/options were presented to customers. The image on the top left is a screen shot of the app when saving preferences via a mobile device. The image on the top right is a screen shot of the app when placing an order via mobile device. The bottom image is a screen shot of a completed order as presented on a computer.

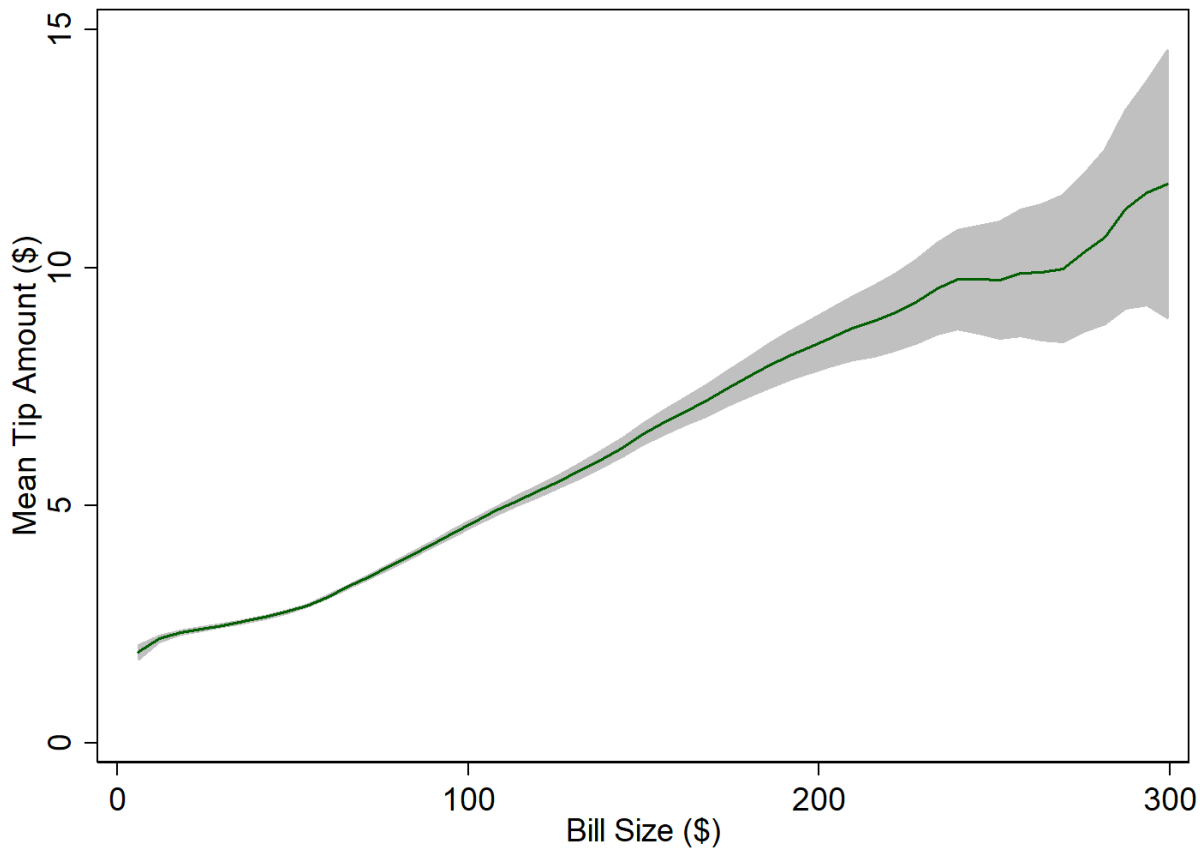


Figure 2. The relationship between tip amount and bill size (for the 99.25% of transactions with bill size below \$300). The figure displays a local polynomial smooth of tip amount on bill size; the confidence intervals displayed are not adjusted for the presence of multiple transactions per customer.

Table 1. Labels and definitions of variables used in this study.

Customer Id = A unique random identifier given to each customer and used to assign customers to tip suggestion conditions.
Tip Suggestion Condition = A categorical variable indicating which set of tip suggestions the invoiced customer received – A = \$2, \$4, \$6, B = \$3, \$4, \$5, C = \$3.95, \$4, \$4.05, D = \$3.99, \$5.99, \$7.99, E = \$4, \$6, \$8, F = \$5, \$8, \$10, G = 5%, 10%, 15%, H = 10%, 15%, 20%, I = 12%, 15%, 18%, J = 15%, 18%, 20%, or K = 15%, 20%, 25%.
Ask Amount = A variable reflecting the average percentage tip amounts recommended in three otherwise comparable tip recommendation conditions – condition G (coded as 10), condition H (coded as 15) and condition K (coded as 20). Other conditions were coded as a missing value on this variable.
G%vsA\$ = A measure of dollar vs. percentage recommendation coded as 0 when condition was A (\$2, \$4, \$6) and coded as 1 when condition was G (5%, 10%, 15%). Other conditions were coded as a missing value on this variable.
H%vsE\$ = A measure of dollar vs. percentage recommendation coded as 0 when condition was E (\$4, \$6, \$8) and coded as 1 when condition was H (10%, 15%, 20%). Other conditions were coded as a missing value on this variable.
Invoice Date = The date the customer was first invoiced for a completed order – typically right after delivery of the cleaned clothes. Recoded as days before and after (centered on) September 15, 2015.
Bill Size = The dollar and cent amount (before coupons or other discounts) of the order.
Bill Size Variability = Absolute deviation of bill size from the mean in that customers' tip suggestion condition.
Tip Likelihood = A binomial variable indicating whether the customer left a tip for the order or not.
Tip Amount = Dollar and cent amount left as a tip (including zero) for the order.
Size of Non-Zero Tip = Dollar and cent amount left as a tip when a tip was left (excluding tips of zero) for the order.
Tip Variability = Absolute deviation of tip amount from the mean in that customers' tip suggestion condition.
Satisfaction Rating Provided = A binomial variable indicating whether the customer left a rating for the order or not.
Satisfaction Rating = Number of stars (on a five star scale) the customer gave the invoiced service transaction.
Satisfaction Variability = Absolute deviation of satisfaction rating from the mean in that customers' tip suggestion condition.
Customer Patronage Frequency = Number of times the customer patronized Washio after the app change.

Patronage Frequency Variability = Absolute deviation of Customer Re-patronage from the mean in that customer tip suggestion condition.
Saved Tip Preference = A binomial variable indicating whether the customer had a saved tip preference or not.
Default Tip = A binomial variable indicating whether the customer left a default tip or not.
Order Number = A variable reflecting the ordinal position of the order in the set of all orders by that customer made after the app change and arranged by date – e.g., 1 = customer’s first order, 2 = customer’s second order, etc...
Median Income = median income of the zip code where the order was picked up and delivered.

Table 2. Descriptive statistics for the key quantitative variables used in this study.

Variable	N	Minimum	Maximum	Mean	Std. Deviation
Ask Amount	26149	10	20	14.63	4.04
G% vs A\$	17767	0 = A\$	1 = G%	0.54	.50
H% vs E\$	16225	0 = E\$	1 = H%	0.55	.50
Invoice Date (days after 9/15/15)	94571	0	359	146.16	99.29
Bill Size	94571	5.99	1365.60	70.58	42.67
Bill Size Variability	94571	.00	1295.00	28.78	31.50
Percent Tip	94571	.00	133.56	5.46	5.95
Tip Likelihood	94571	0 = no	1 = yes	.59	.49
Size of Non-Zero Tip	55495	.01	121.52	6.14	4.94
Tip Amount	94571	.00	121.52	3.61	4.85
Tip Variability	94571	.00	116.22	3.24	3.39
Satisfaction Rating	74348	1	5	4.53	.97
Satisfaction Variability	74348	.41	3.59	.69	.68
Customer Re-patronage frequency	94571	0 = no	1 = yes	.50	.50
Customer Patronage Frequency	24637	1	69	3.84	5.11
Patronage Frequency Variability	24637	.04	65.16	3.28	3.92
Saved Tip Preference	94571	0 = no	1 = yes	.49	.50
Default Tip	94571	0 = no	1 = yes	.49	.50
Order Number	94571	1	69	5.82	6.55
Median Income	89513	\$19,887	\$172,570	\$81,429.13	\$28,497.09

Table 3. Pre-treatment means (and standard deviations) of customer satisfaction, re-patronage and spending for orders before September 15, 2015, by tip suggestion condition.

Tip Suggestion Condition	Satisfaction Rating	Customer Patronage Frequency ^a	Bill Size
A: \$2, \$4, \$6	4.48 (.98)	3.76 (5.55)	54.13 (34.24)
B: \$3, \$4, \$5	4.51 (.97)	3.74 (5.28)	54.40 (47.66)
C: \$3.95, \$4, \$4.05	4.48 (.99)	3.84 (5.64)	54.72 (36.38)
D: \$3.99, \$5.99, \$7.99	4.51 (.96)	3.69 (5.39)	54.63 (38.29)
E: \$4, \$6, \$8	4.50 (.94)	3.59 (5.13)	54.33 (35.56)
F: \$5, \$8, \$10	4.51 (.95)	3.63 (5.14)	55.26 (37.30)
G: 5%, 10%, 15%	4.49 (.97)	3.75 (5.51)	54.75 (38.35)
H: 10%, 15%, 20%	4.52 (.95)	3.78 (5.61)	54.96 (50.44)
I: 12%, 15%, 18%	4.46 (.98)	3.75 (5.40)	54.22 (36.23)
J: 15%, 18%, 20%	4.50 (.96)	3.57 (5.02)	54.24 (36.61)
K: 15%, 20%, 25%	4.50 (.95)	3.71 (5.58)	56.35 (39.67)
Omnibus test (df)	F(10, 27,750) ^b	F(10, 39,102) ^c	F(10, 39,122) ^b
Value of Test Statistic	1.00	0.93	0.58
R ²	.0003	.0002	.0002
N orders/customers	97,078/27,750	NA/39,113	145,324/39,113

^a Number of times the customer patronized Washio before the app change.

^b Robust error terms clustered within customer.

^c Robust error terms.

* $p < .05$, ** $p < .01$, *** $p < .001$, **** $p < .0001$

Table 4. Means (and standard deviations) of measures of tipping behavior by tip suggestion condition.

Tip Suggestion Condition	Size of Non-Zero Tip	Tip Likelihood	Tip Amount	Tip Variability	Has Saved Tip Preference	Left Default Tip
A: \$2, \$4, \$6	\$3.57 ^{BEG} (1.81)	.64 ^{beG}	\$2.28 ^{bEG} (2.24)	\$1.81 ^{bEG} (1.32)	.52 ^{bEg}	.61 ^{BEG}
B: \$3, \$4, \$5	\$3.79 ^A (1.33)	.61 ^a	\$2.30 ^a (2.12)	\$1.84 ^a (1.05)	.49 ^a	.56 ^A
C: \$3.95, \$4, \$4.05	\$4.04 (1.38)	.60	\$2.43 (2.25)	\$1.98 (1.07)	.50	.49
D: \$3.99, \$5.99, \$7.99	\$4.38 ^E (1.49)	.60 ^e	\$2.62 ^e (2.44)	\$2.22 ^E (1.01)	.49 ^e	.47 ^E
E: \$4, \$6, \$8	\$4.62 ^{ADH} (1.61)	.60 ^{adh}	\$2.79 ^{AdH} (2.58)	\$2.31 ^{ADH} (1.15)	.47 ^{Adh}	.52 ^{ADH}
F: \$5, \$8, \$10	\$5.43 (2.23)	.57	\$3.12 (3.17)	\$2.75 (1.58)	.45	.44
G: 5%, 10%, 15%	\$6.66 ^{AHK} (5.03)	.59 ^{AehK}	\$3.91 ^{AHK} (5.06)	\$3.63 ^{AHK} (3.53)	.51 ^{ahK}	.52 ^{AhK}
H: 10%, 15%, 20%	\$8.07 ^{EgIK} (6.00)	.58 ^{egiK}	\$4.70 ^{EgIK} (6.07)	\$4.27 ^{EgIK} (4.31)	.50 ^{egiK}	.48 ^{EgIK}
I: 12%, 15%, 18%	\$8.52 ^h (5.80)	.59 ^h	\$5.05 ^h (6.12)	\$4.45 ^h (4.20)	.51 ^h	.43 ^H
J: 15%, 18%, 20%	\$9.62 (7.40)	.56	\$5.36 (7.30)	\$5.13 (5.20)	.47	.39
K: 15%, 20%, 25%	\$11.23 ^{GH} (7.62)	.50 ^{GH}	\$5.63 ^{GH} (7.79)	\$5.80 ^{GH} (5.19)	.44 ^{GH}	.43 ^{GH}
Omnibus test (df)	F(10, 16,530)	Wald χ^2 (10)	F(10, 24,636)	F(10, 24,636)	Wald χ^2 (10)	Wald χ^2 (10)
Value of Test Statistic	271.72****	56.41****	99.27****	307.57****	24.56**	192.24****
R ² /Pseudo R ²	.24	.003	.06	.15	.0016	.0098
N orders/customers	55,495/16,531	94,571/24,637	94,571/24,637	94,571/24,637	94,571/24,637	94,571/24,637

Notes: Alphabetic superscripts indicate the conditions each statistic was compared with; CAPITALIZED superscripts mark comparisons reliable at the .05 level unadjusted for multiple comparisons; and underlined superscripts mark comparisons reliable at the .05 level adjusted for multiple comparisons using the method of List, Shaikh and Xu (2016). Analyses adjusting for multiple comparisons were based on one randomly selected observation per customer and included all paired-comparisons, but tests were performed separately for the following groups of outcomes: {size of non-zero tip, saved preference, default tip} and {tip likelihood, tip amount, tip variability}. For the omnibus tests, we report F statistics for regressions estimated using OLS, and χ^2 statistics from Wald tests for regressions estimated using binomial logistic regression. Robust standard errors for omnibus regressions are adjusted for clustering at the customer level. ** p < .01, ****p < .0001

Table 5. Tip amount vs. ask amount and interactions with other variables.

	(1)	(2)	(3)	(4)	(5)	(6)
	Tip Amount	Tip Amount	Tip Amount	Tip Amount	Tip Amount	Tip Amount
Ask amount	0.172*** (0.0261)	-0.114 (0.0649)	0.0538** (0.0208)	0.0486 (0.0734)	0.154* (0.0744)	0.0905* (0.0396)
Bill size		0.0106 (0.0153)				
Ask × Bill size		0.00394*** (0.00104)				
Saved preference			0.412 (0.585)			
Ask × Saved preference			0.323*** (0.0423)			
Median income (\$000)				-0.0263* (0.0123)		
Ask × Median income				0.00159 (0.000876)		
Satisfaction rating					0.303 (0.224)	
Ask × Satisfaction rating					0.00617 (0.0161)	
Order number						-0.239 (0.176)
Ask × Order number						0.0481*** (0.0133)
<i>R</i> ²	0.012	0.233	0.186	0.014	0.016	0.028
Customers	6714	6714	6714	6509	4944	2054
Observations	26149	26149	26149	24793	20537	8216

Notes: The table reports coefficients from OLS regressions of tip amount on ask amount and a number of interactions, with the sample restricted to only those customers in treatment conditions G (5%, 10%, 15%), H (10%, 15%, 20%), and K (15%, 20%, 25%). *Ask amount* is defined as the suggested percentage of the middle option, so it is equal to 10 or 15 or 20. Median income represents the median income of the customer's zip code; and saved preference is an indicator for whether the customer chose to save their tip preference. In order to control for compositional effects, the sample in column (6) is restricted to only those customers with 4 or more orders *and* only the first 4 orders of those customers; thus order number is an integer ranging from 1 to 4. Robust standard errors in parentheses, adjusted for clustering at the customer level.

* $p < .05$, ** $p < .01$, *** $p < .001$

Table 6. The effect of recommendation format (percentage vs. dollar) on tip amount.

	All bill sizes		All bill sizes		Bill size < \$40		\$39.50 < Bill size < \$40.50	
	(1) Tip Amount	(2) Tip Amount	(3) Tip Amount	(4) Tip Amount	(5) Tip Amount	(6) Tip Amount	(7) Tip Amount	(8) Tip Amount
G% vs A\$	1.635*** (0.151)		-1.359** (0.414)		0.116 (0.133)		0.846** (0.277)	
H% vs E\$		1.912*** (0.164)		-2.171*** (0.384)		0.255 (0.152)		0.264 (0.356)
Bill size			0.00970*** (0.00114)	0.00785*** (0.00140)				
G% vs A\$ × Bill size			0.0419*** (0.00661)					
H% vs E\$ × Bill size				0.0582*** (0.00611)				
R^2	0.040	0.038	0.201	0.217	0.001	0.003	0.047	0.003
Customers	4423	4307	4423	4307	1692	1705	279	234
Observations	17767	16225	17767	16225	3473	3480	369	299

Notes: The table reports the coefficients from OLS regressions of tip amount on an indicator variable for whether the tip recommendation was provided in percentage terms. G% vs A\$ is equal to 1 when the suggestion condition is G (percentage) and 0 when it is A (dollar amounts), and H% vs E\$ is defined analogously. Columns (3) and (4) include interactions between the treatment indicators and bill size. Each sample is restricted to only the two suggestion conditions being compared (i.e., G and A, or H and E). Robust standard errors in parentheses, adjusted for clustering at the customer level. * $p < .05$, ** $p < .01$, *** $p < .001$

Table 7. Means (and standard deviations) of various measures by tip suggestion condition.

Tip Suggestion Condition	Satisfaction Rating	Satisfaction Variability	Customer Patronage Frequency	Patronage Frequency Variability	Bill Size	Bill Size Variability	Satisfaction Rating Provided
A: \$2, \$4, \$6	4.53 ^{beg} (.96)	.69 ^{beG} (.67)	3.96 ^{beg} (5.21)	3.36 ^{bEg} (3.98)	71.36 ^{beg} (42.09)	28.86 ^{beg} (30.63)	.78 ^{beg}
B: \$3, \$4, \$5	4.52 ^a (.97)	.70 ^a (.67)	3.82 ^a (5.06)	3.23 ^a (3.89)	69.44 ^a (39.50)	27.51 ^a (28.34)	.79 ^a
C: \$3.95, \$4, \$4.05	4.52 (.97)	.70 (.67)	3.84 (5.22)	3.28 (4.06)	69.96 (40.93)	28.29 (29.58)	.78
D: \$3.99, \$5.99, \$7.99	4.52 ^e (1.01)	.71 ^e (.72)	3.83 ^e (5.10)	3.29 ^e (3.90)	71.13 ^e (42.42)	28.66 ^e (31.28)	.80 ^e
E: \$4, \$6, \$8	4.54 ^{adh} (.97)	.69 ^{adh} (.68)	3.72 ^{adh} (4.76)	3.12 ^{A^{adh}} (3.60)	69.42 ^{adh} (42.94)	28.74 ^{adh} (31.90)	.78 ^{adh}
F: \$5, \$8, \$10	4.54 (.94)	.68 (.65)	3.82 (5.20)	3.25 (4.05)	71.36 (44.61)	30.00 (33.02)	.79
G: 5%, 10%, 15%	4.59 ^{aHK} (.93)	.64 ^{AHK} (.67)	4.07 ^{ahk} (5.65)	3.57 ^{aHK} (4.37)	71.47 ^{ahk} (42.91)	29.52 ^{ahk} (31.15)	.78 ^{ahk}
H: 10%, 15%, 20%	4.53 ^{eGik} (.98)	.70 ^{eGik} (.68)	3.81 ^{egik} (4.99)	3.29 ^{eGik} (3.75)	70.07 ^{egik} (42.31)	28.54 ^{egik} (31.24)	.80 ^{egik}
I: 12%, 15%, 18%	4.53 ^h (.98)	.70 ^h (.68)	3.88 ^h (4.96)	3.27 ^h (3.74)	69.78 ^h (42.01)	27.81 ^h (31.48)	.80 ^h
J: 15%, 18%, 20%	4.52 (.95)	.70 (.65)	3.63 (4.67)	3.02 (3.56)	70.60 (48.05)	29.56 (37.88)	.77
K: 15%, 20%, 25%	4.51 ^{Gh} (.98)	.71 ^{Gh} (.67)	3.80 ^{gh} (5.18)	3.31 ^{Gh} (3.98)	72.05 ^{gh} (42.55)	29.39 ^{gh} (30.78)	.78 ^{gh}
Omnibus test (df)	F(10, 18135)	F(10, 18135)	Wald χ^2 (10)	Wald χ^2 (10)	F(10, 24636)	F(10, 24636)	Wald χ^2 (10)
Value of Test Statistic	1.21	3.04***	10.62	26.27**	0.72	1.28	10.75
R ² /Pseudo R ²	.0004	.0008	.0002	.0005	.0004	.0006	.0006
N orders/customers	74,348/18,136	74,348/18,136	NA/24637	NA/24637	94571/24637	94571/24637	94,571/24,637

Notes: Alphabetic superscripts indicate the conditions each statistic was compared with; CAPITALIZED superscripts mark comparisons reliable at the .05 level unadjusted for multiple comparisons; and underlined superscripts mark comparisons reliable at the .05 level adjusted for multiple comparisons using the method of List, Shaikh and Xu (2016). Analyses adjusting for multiple comparisons were based on one randomly-selected observation per customer and included all paired-comparisons, and tests were performed separately for the following groups of outcomes: {satisfaction rating, satisfaction variability}, {patronage frequency, bill size, rating provided}, and {frequency variability, bill size variability}. F statistics are reported for OLS regressions, and Wald χ^2 statistics for regressions estimated using negative binomial regression (patronage frequency and variability) or logistic regression (rating provided). Robust standard errors for omnibus regressions are adjusted for clustering at the customer level. ** p < .01, *** p < .001, ****p < .0001

Table 8. Reliability (with and without adjustment for multiple comparisons) and direction of effect for key paired comparisons.

	Tip Likelihood	Size of Non-Zero Tip	Tip Amount	Tip Variability	Satisfaction	Satisfaction Variability	Patronage Frequency	Patronage Frequency Variability	Bill Size	Bill Size Variability	Satisfaction Rating Provided	Has Saved Tip Preference	Left Default Tip	
<u>Suggest Large vs. Small Amounts</u>														
E vs. A	+*	+	+	+	n.s.	n.s.	n.s.	(-)	n.s.	n.s.	n.s.	-	-	
H vs. G	n.s.	+	+	+	(-)	(+)	n.s.	(-)	n.s.	n.s.	n.s.	n.s.	n.s.	
K vs. H	-	+	(+)	+	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	-	-	
K vs. G	-	+	+	+	(-)	(+)	n.s.	(-)	n.s.	n.s.	n.s.	-	-	
<u>Suggest Wide vs. Narrow Range of Amounts</u>														
A vs. B	n.s.	+	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	(+)	
H vs. I	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	+	
<u>Suggest Round vs. Non-Round Amounts</u>														
E vs. D	n.s.	(+)	n.s.	(+)	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	(+)	
<u>Suggest Percentage vs. Dollar Amounts</u>														
G vs. A	(-)	+	+	+	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	-	
H vs. E	n.s.	+	+	+	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	+*	(-)	

Notes: Conditions used in comparisons were: A = \$2, \$4, \$6, B = \$3, \$4, \$5, E = \$4, \$6, \$8, D = \$3.99, \$5.99, \$7.99, G = 5%, 10%, 15%, H = 10%, 15%, 20%, I = 12%, 15%, 18% and K = 15%, 20%, 25%. Signs mark larger (+), smaller (-), or non-significantly different (n.s.) means in the condition on the left than in the condition on the right. All alphas were .05. Effects in parentheses were reliable in non-adjusted analysis of the entire sample, but were unreliable in analysis of a reduced sample (one randomly selected observation per customer) that adjusted for multiple comparisons using List, Shaikh and Xu's (2016) method and code. Effects with asterisks were not significant in non-adjusted analysis of the entire sample, but were significant in the analysis of a reduced sample that adjusted for multiple comparisons. Effects with neither parentheses or asterisks were reliable in both adjusted and non-adjusted analyses.

Table 9. Post study probabilities of ask-size effects and ask-size null-effects given various prior probabilities of those effects.

Prior Probability of Effect (π)	Ask-Size Effects on Tip Amount			
	$E - A \geq \$0.51$	$H - G \geq \$0.79$	$K - H \geq \$0.93$	$K - G \geq \$1.72$
.01	.17	.17	.17	.17
.05	.51	.51	.51	.51
.10	.69	.69	.69	.69
.25	.87	.87	.87	.87
.50	.95	.95	.95	.95

Prior Probability of Null-Effect (π')	Ask-Size Null-Effects on Satisfaction Ratings			
	$ E - A < 0.1$	$ H - G < 0.1$	$ K - H < 0.1$	$ K - G < 0.1$
.01	.04	.03	.03	.03
.05	.16	.14	.12	.13
.10	.29	.26	.23	.24
.25	.55	.51	.47	.48
.50	.78	.76	.73	.74

Notes: Conditions used in comparisons of ask-size were: A = \$2, \$4, \$6, E = \$4, \$6, \$8, G = 5%, 10%, 15%, H = 10%, 15%, 20%, and K = 15%, 20%, 25%. The post study probabilities of our observed ask-size effects being true is $(1-\beta)\pi/((1-\beta)\pi + \alpha(1-\pi))$, where $1-\beta$ is power (post-hoc power from current study contrast in this case), π is the prior probability of effect truth, and α is the alpha level (.05 in this case). The post study probabilities of our observed ask-size null-effects being true is $(\pi'(1-\alpha)/((\pi'(1-\alpha) + \beta(1-\pi')))$, where true null-effects are defined as absolute differences between condition means that are less than the smallest meaningful amount (.1 out of a 5-point scale in this case), π' is the prior probability of a null-effect being true, β is the chance of getting a non-significant result if the smallest meaningful effect is true, and α is the alpha level (.05 in this case).