W241 Tipping Project

Ben Attix, Chris Beecroft, Brock Kowalchuk, Matthew Burke

Does all-inclusive (tax + tip) pricing at restaurants impact the customer perception of value?

Introduction

Tipping varies around the world from countries such as Australia where tipping is not practiced, to countries such as Germany where rounding the bill up to the nearest Euro is the norm. In the US, the expectation for tipping has slowly climbed from 10 to 15 to 18 to 20 percent over the past 100 years [1]. Often a mandatory tip is added for large groups and some high-end restaurants automatically add an 18% experience or service charge. While the idea of giving a small gratuity for services is old, the roots of modern US tipping practices go back to the post-Civil War era. According to Saru Jayaraman of the Food Labor Research Center, UC Berkeley, tipping started as a way for train operators to avoid paying the newly-freed slaves and still have them earn an income [2]. This practice continues today with the alternate minimum wage that restaurant owners are allowed to pay their staff on the premise that tips will make up the difference.

This two-tiered system "traps many low-wage tipped restaurant workers in conditions of economic and social vulnerability" [3]. A continually rising 'accepted' tip rate is not sustainable and the movement to create all-inclusive pricing is an attempt at addressing rising expected tips and low wages in the food industry. If Americans are willing to bear the burden of the true cost of restaurants, then perhaps we can achieve better pay equality. It is this concern of fairness and social responsibility that drives our research question.

A small number of restaurants in cities such as San Francisco, Berkeley, and New York are attempting to address this by creating an all-inclusive pricing model. Danny Meyer, a respected New York restaurateur, made a splash a few years ago when he announced he would move all of his restaurants to this pricing structure [4]. While he has since retreated from using this model at some of his restaurants, the attempt to change the system still exists. The question we seek to answer is the following: if a patron knows an establishment is tip-less, will they accept the higher price and feel that the additional cost is reasonable. Our hypothesis is that they would be less satisfied with meal prices with this model.

---

[1] Tuttle, B. (2014-09-18). 15 Things You Didn't Know About Tipping. Time Magazine. Retrieved from http://time.com/money/3394185/tipping-myths-realities-history/ on 2017-05-02.

[2] Ferdman, R. A. (2016-02-18). I dare you to read this and still feel good about tipping. The Washington Post. Retrieved from https://www.washingtonpost.com/news/wonk/wp/2016/02/18/i-dare-you-to-read-this-and-still-feel-ok-about-tipping-in-the-united-states/?utm_term=.a694c2615e80 on 2017-05-02.

[3] Food Labor Research Center (2015-12). Working Below The Line. UC Berkeley. Retrieved from http://food.berkeley.edu/wp-content/uploads/2015/07/WorkingBelowTheLine_F2.pdf on 2017-05-02.

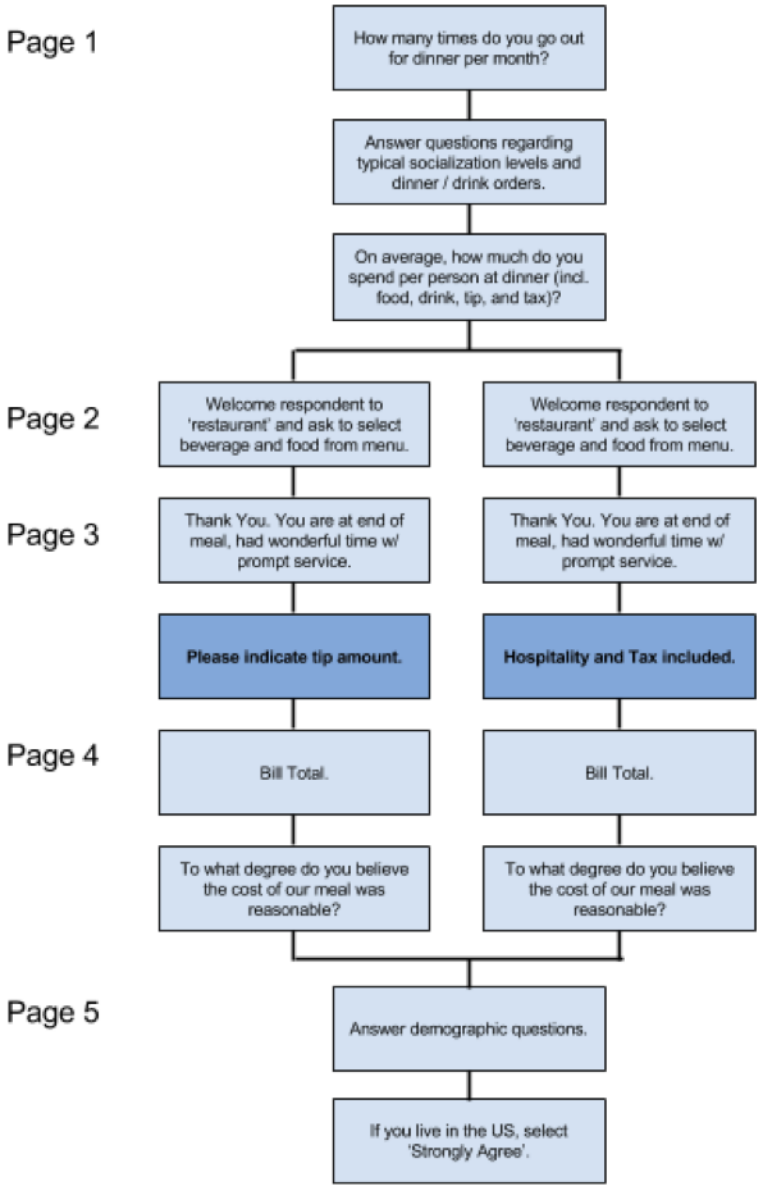
[4] Sutton, R. (2015-10-14). Danny Meyer is eliminating all tipping at his restaurants and significantly raising prices to make up the difference, a move that will raise wages, save the hospitality industry, and forever change how diners dine. New York Eater. Retrieved from https://ny.eater.com/2015/10/14/9517747/danny-meyer-no-tipping-restaurants on 2017-05-02.

Experiment

At the start of our experimental design phase, we researched Bay Area restaurants that were offering all-inclusive pricing and reached out to a few restaurateurs we knew. Asking a restaurant to randomly select patrons to give them separately priced menus or to switch up nights that they used one system or another seemed problematic. We opted instead for a restaurant experience "thought experiment" design that would use a Qualtrics survey (see appendix) driven by subjects from Amazon Mechanical Turk.

Qualtrics Survey

In brief, the survey design was to ask a few questions about the subject's dining habits (appendix, page 1), then randomly assign the subject to the control or treatment group, where they would be presented with a menu of drink and food options (page 2). They would then be presented with a bill and asked for a tip if they were in the control group (page 3). They were shown the total bill amount and were asked if they felt the cost was reasonable (page 4). They were then shown a final set of questions (page 5), and then the completion page with their Mechanical Turk code (page 6). Below is the flow of the survey with control being on the left and treatment on the right. The survey page numbers are called out on the left side for reference.



Page 1 contained the basic covariates we were interested in for evaluating the survey results: Number of times eating out, who they normally dined with (multi-choice), which courses they normally ordered (multi-choice), what drinks they normally ordered (multi-choice), and their average spend per person when dining with a significant other. (As an aside, in the appendix, the bottom right footer says 1/2; page 2 was blank and has been omitted.)

Page 2 Control, Page 2 Treatment: We started with the text, "Welcome to our restaurant, please be seated. Here is a menu, please look it over and select what you would like. I will be back shortly." For the food section of the menu we had several concerns. One was that we did not want to provide a name for the restaurant as we felt this could influence the perception of pricing. Certain words such as diner or café could be interpreted differently in different parts of the country, so we opted for no restaurant name.

We opted for a world cuisine loosely based on California- and New York- based restaurants, as we felt that a regional cuisine such as French or Italian would put people into a particular mind-set. We wanted a mix of vegetarian and meat options and wanted to include an item that would be more familiar to North African/Mediterranean subjects. For inspiration we reviewed the a la cart menus from Berkeley's Chez Pannise Cafe. While we wanted to have a consistent feel in the language, we stayed away from words that would be unfamiliar to subjects while trying portray a higher level of dining and variety that would appeal to a broader audience.

Pricing was another issue that we worked through. Prices are widely varied in the US depending on whether you are in an urban area or a rural part of the country. We decided that we would focus on the urban areas. While the prices on the control menu are inexpensive compared to Bay Area or New York dining, they are not overly low. For the treatment menu, we chose a price that was about 30% higher on average (between 27% and 32%). This number was selected on the idea that this would incorporate a 7% tax and an 18% gratuity with an a few additional percentage points of surcharge. Also of note is that the dinner items were randomly shuffled for each survey.

We opted to add a beverage section, including a free water option, to give subjects the option to select a no-cost item, rather than requiring a beverage purchase. For alcoholic beverages, prices in the treatment group were $1 higher than in the control group, with the exception of the cocktail, which was priced $2 higher, given that mixed drinks are more labor intensive. These amounts were chosen to reflect typical tip amounts in urban areas.

We ended the treatment page with the text "Hospitality and tax included." This was loosely based on text that Danny Meyer used in his New York Restaurants.

Page 3 Control, Page 3 Treatment: For page 3, the subjects were told that they had a wonderful time and service was prompt. They were presented with an itemized bill. The control group was then asked to provide a tip amount while the treatment group was again reminded that hospitality and tax were included. (In the sample survey in the appendix, we show the same menu items for control and treatment, plus and a $7 tip on the control page.)

Page 4 does not differ between Control and Treatment. The two are broken out here to show the cost difference between the flows.

Page 5: After our initial pilot study, we found that some participants were clicking through the survey in 30 seconds or less. At the recommendation of the Qualtrics user guide, we added the question, "If you live in the U.S. select strongly agree". Since we set location restrictions and only included people in the U.S. that means everyone should have selected "strongly agree". For people who did not choose "strongly agree", we can conclude that they were not paying attention to the survey and therefore we excluded their responses. We also added questions on gender, age, and average tip amount in the hope of making this question less obvious to subjects that clicked through the survey.

Page 6 is the end of the survey.

Since the cost of food and restaurants is quite varied across the United States, we also wanted to limit participation to only a few urban areas.

Amazon Mechanical Turk

With Amazon Mechanical Turk, we set up a task with the description: "Let us know your experience and preference at restaurants (3-5 minute survey)." While Qualtrics estimated the survey should take 6 minutes, the mean survey time was just over 2 minutes. The task was offered at $ 0.30, which would place a 2-minute subject above the national minimum wage.

For our pilot studies we were able to quickly get the requested number of subjects, but for the actual survey, the $ 0.30 was on the low side, and it took twenty days before we decided to end the survey. We had planned to stop after 330 responses (300 responses and a 10% overage for the inevitable re-takers from the pilot surveys). The survey started on April 5th and only garnered three responses after three days. We realized that we had one of the restrictions backwards (limit to users who have 0 tasks completed, switched to users who have more than 0 tasks completed). We restarted the survey on the 8th before finally deciding on Thursday April 20th to end the survey on Sunday April 23rd. We did contemplate raising the price but felt that this would affect the results of the survey. From these two sets we received 263 responses.

Finally we set up Mechanical Turk to limit data to California, New York, Pennsylvania, and Washington D.C. Unfortunately Mechanical Turk does not provide a more fine-grained selection of areas, so we targeted high-population states as a proxy for urban areas. The validation of the region selection data was done as one of our final pilots where we collected 30 subjects on April 3rd to validate that the region selections were working correctly. During validation of the area restriction pilot, we examined the latitude and longitude in the data set to verify that subjects were in our target areas. No other data was reviewed. These were included to make up for the unreached number.

Research Design (RXO Grammar)

Once we had established that we wanted to do a simple survey, our options for research design were limited, and we settled on a standard "RXO" model to satisfy our data needs. RXO stands for research, treatment, outcome which directly follows the one-time data collection we had our subjects step through in the survey. First, we collected some covariate information, randomized them into treatment or control, applied the treatment or control menu and measured the outcome variable response in the final stage. While this isn't a particulalry sophisticated or unique model, it satisfies the purpose of running a between-subjects design on an outcome variable that could be affected by viewing it more than once.

Group	Randomization	Treatment	Outcome
Treatment	R	X	O
Control	R	–	O

Our main concern with the treatment step of providing the subject with a bill with standard or all-inclusive pricing was the selection of the meal options. If we limited the menu to a single item, then taste or cost trends in a particular region may bias the results regardless of what the subject thinks about tipping, as well as not accurately simulating a real restaurant environment with a variety of options. Thus we made the choice to simulate a real menu as closely as possible with items across a range of pricing and ingredient options with the hope that at least one or two meals

would appeal to the subjects' preferences and budget. One assumption this decision necessitated was that the subjects choose menu items using the same thought process as they would in store, considering cost as a factor. If we don't have this assumption, this study is relatively nonsensical as subjects may choose a meal outside of their price range and complain at its unreasonableness. We would not expect this type of behavior from an in-person interaction, and so for this experiment, we also expect the subjects to behave as they normally would.

Experimental Materials & Randomization Engineering

The ideal experiment to determine whether all-inclusive pricing impacted the satisfaction levels of restaurant diners would have included selecting a random sample of restaurants across an urban environment (i.e. New York or San Francisco) and within each restaurant, randomly selecting individuals to be treated to the all-inclusive rate or the standard rate. This would have allowed a live experiment where we may have been able to cut across multiple covariates to determine more granular reasoning for a potential outcome. However, given the lack of resources and restaurateur connections, the next best option was to somewhat replicate the experience of eating in a restaurant via online surveys.

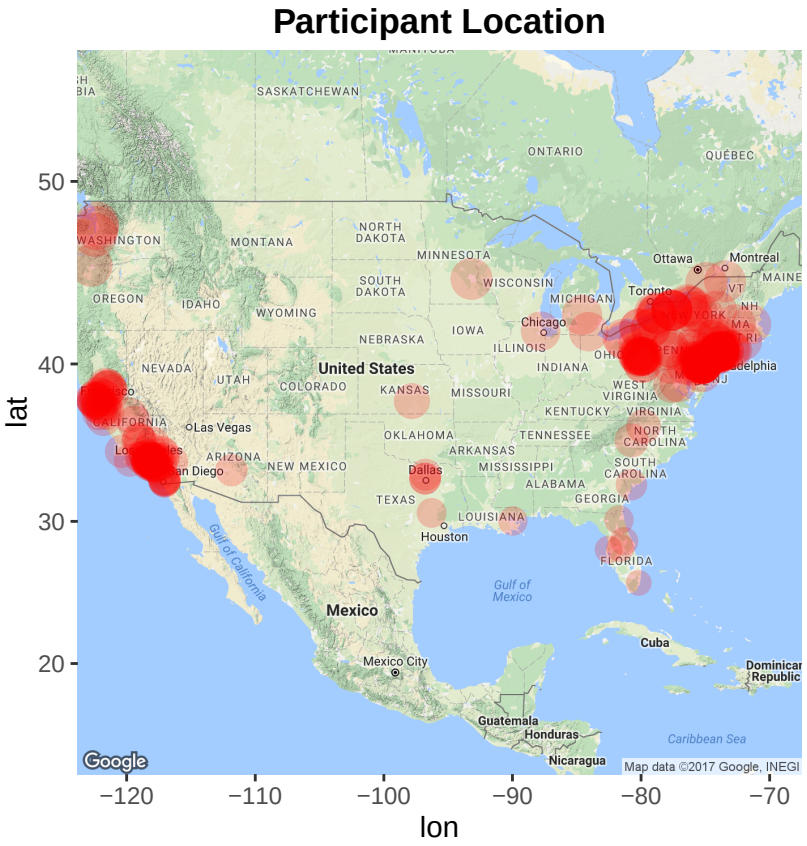
Prior to the final experiment, the group conducted several preliminary surveys to learn the nuances around Amazon's Mechanical Turk and Qualtrics. Through these early surveys the team was able to improve the survey flow and construction to provide a more robust outcome including:

Mitigating risk of repeat survey participants Restricting respondent's locations to California, New York, Pennsylvania, and DC Implementing a Compliance Check near end of survey Randomizing so equal number of treatment and control respondents flow through survey

While the group did not succeed in preventing repeat responders from initiating a second survey, the IP Address of each user was captured from Mechanical Turk and this was used to filter all results from duplicate individuals from the final analysis. Furthermore, the preliminary surveys yielded responses from individuals primarily from outside the United States. The team subsequently decided to isolate responses from four states and districts within the country that have large metropolitan areas. The plot below shows the high concentrated of respondents in the desired regions. The responses captured from outside these regions could be due to extraneous factors (e.g. Turks vacationing or moving locations but not having their profile updated).

```
## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=United+States&zoom=4&siz

## Information from URL : http://maps.googleapis.com/maps/api/geocode/json?address=United%20Sta
```



A significant benefit online is the ease at which randomization can be executed, however, the difficulty is sometimes identifying which unobserved characteristics could potentially dilute the academic strength of any experimental results. This experiment had one point of randomization - where individuals in the treatment group received a 'restaurant bill' having the hospitality (tip) and tax included and another receiving a more traditional bill where tax was added after the meal and the individual could pencil in a tip.

The key to making randomization a success was to ensure the survey had roughly the same number of people in the treatment group as in the control group. A randomization process was established within Qualtrics where two blocks of questions were created for the treatment and control group, respectively. The group used a 'randomizer' function to randomly present one of the two blocks evenly over the course of the experiment, resulting in the following outcome.

One nuance of the experiment was the uneven number of individual that ultimately ended up in the treatment and control groups. This was the results of not filtering out duplicates and non-compliers until after the experiment was completed. An improved experiment may have an upfront check for duplicates and non-compliance, to mitigate the imbalance between treatment and control.

Measurement of Variables

We gathered a selection of covariate information from our subjects including general demographic data as well as specific questions related to their restaurant-related habits. Below is a table of the different variables we collected along with their types and options.

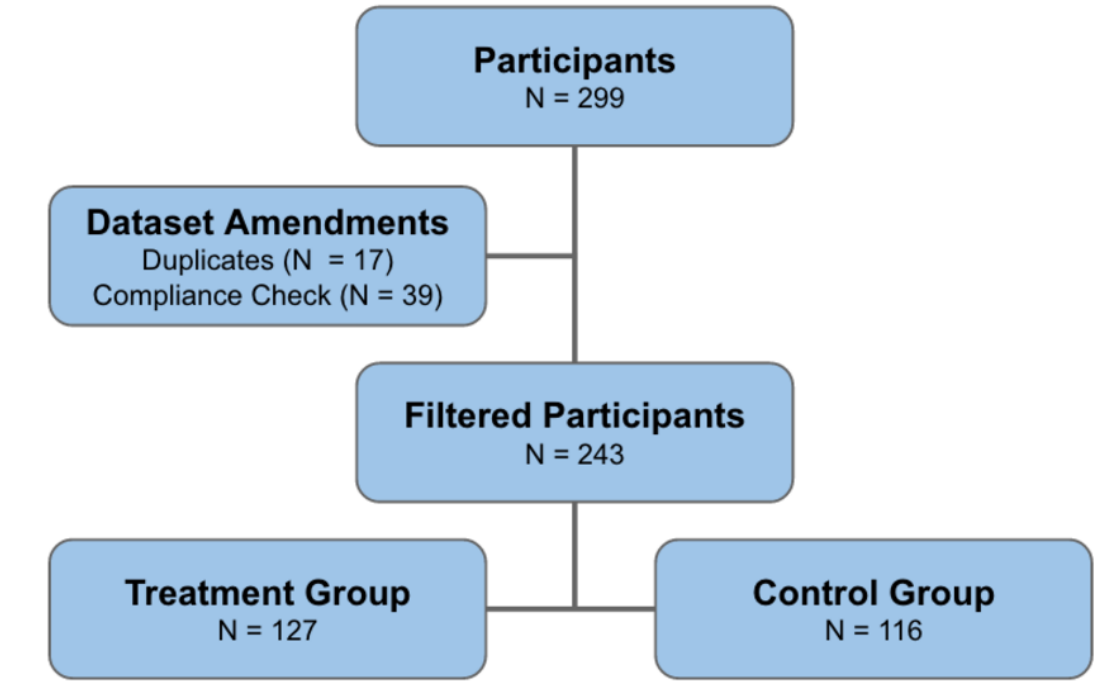


Figure 1: "High-Level Flow Diagram"

Variable	Type	Set
# Times Eat Out Per Month	Numeric	0-30
Who Do You Normally Eat With	Categorical	Friends, Spouse, Significant Other, Family, Alone
Courses Ordered	Categorical	Appetizer, Main Course, Dessert
Drinks Ordered	Categorical	Water, Non-Alcoholic, Wine, Beer or Cider, Mixed Drinks
Average Meal Spend Per Person	Numeric	>0
Gender	Categorical	Male, Female, Prefer Not To Answer
Age	Categorical	Under 18, 18-34, 35-49, 50-64, 65+
Baseline Tip Percentage	Numeric	0-100
Was Meal Cost Is Reasonable	Categorical	Extremely Unreasonable to Extremely Reasonable (1-5)

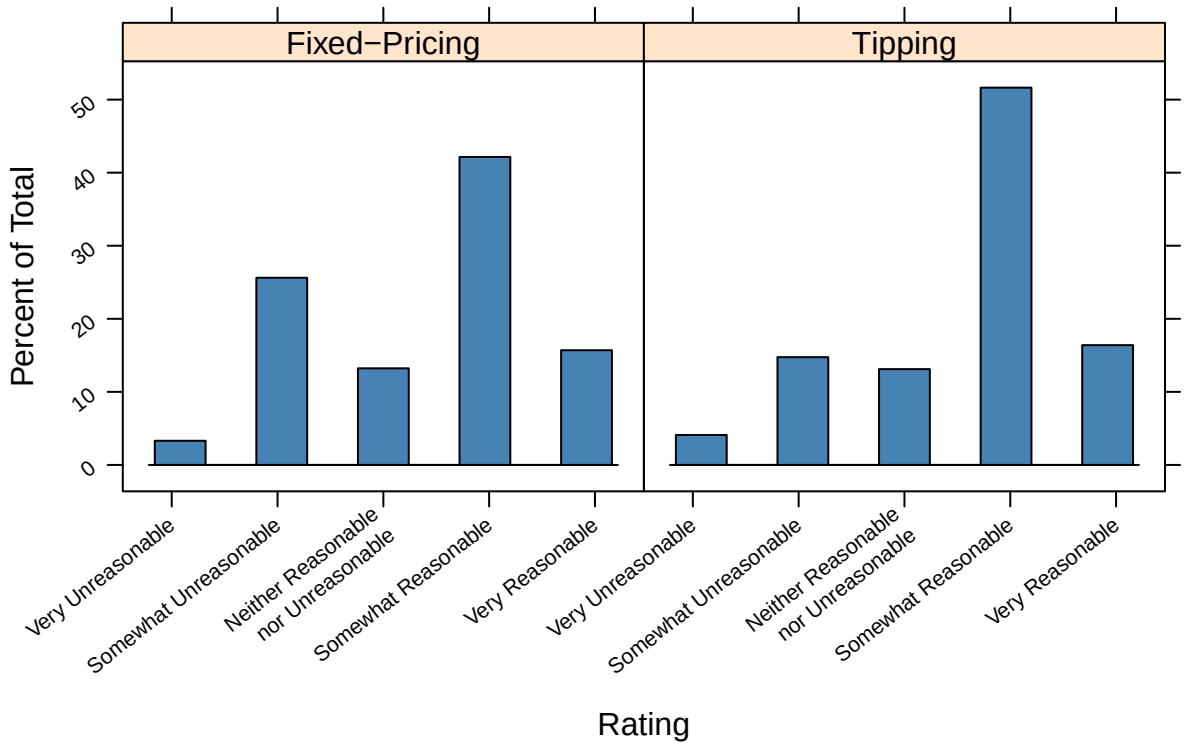
Please note that our outcome variable of the subject's impression of meal price reasonableness was measured using a Likert scale. This is standard practice for questionnaire data collection, and we think it accurately allows us to model their satisfaction with pricing.

Results

Over the span of this course we have used two different methods, linear modeling and randomization inference, to test whether our treatment variable had an effect on our outcome variable. In this section we will present our results using both of these methods as well as a Chi-squared test. Even though we did not discuss Chi-squared tests in this course, we are including it because it is designed for analyzing categorical variables, which we have this experiment.

Figure 1 shows the distribution of responses by treatment group. The average ratings were 3.29 for the treatment group and 3.76 for the control group giving us an average treatment effect of -0.47 before controlling for any other factors.

Figure 1



Linear Modeling

Linear modeling allows us estimate the effect of our treatment variable while also controlling for covariates that may have played a role as well. Table 3 shows three different models of our data. The first model is our "naive" model which uses treatment as the only predictor. The second model includes baseline spending, which is how much the respondent normally spends per person when going out to dinner. Finally, the third model adds in the number of times they normally eat out per month as well as fixed effects for age, gender, who they normally eat with, what courses they normally order, and what drinks they typically order. The third model includes all of the possible covariates that we could have included in our model in order to test whether anything would make a difference in the magnitude or standard error of our treatment estimate. These additional covariates did not do anything for the model and actually made the standard error larger.

Our best model is the second one, because it has the lowest standard error for the treatment variable and it has the highest adjusted R^2. This model shows us an average treatment effect of -0.383 (standard error of 0.129) and a p-value of 0.004. Baseline spending is a significant covariate, but the effect size is only 0.024 or roughly 6% of the size of our treatment effect. Another things to note is that each model shows a statistically significant treatment effect and an ATE between -0.38 and -0.47. Since each model gave such similar results for the ATE, this is a good sign that our treatment truly had an effect and what we're seeing is not due to a poor model.

Table 3:

	Dependent variable:		
	Reasonableness Rating		
	(1)	(2)	(3)
Treatment	−0.467	−0.383	−0.396
	(0.138)	(0.129)	(0.134)
	p = 0.001***	p = 0.004***	p = 0.004***
Baseline Spending		0.024	0.024
		(0.004)	(0.004)
		p = 0.000***	p = 0.00000***
Num Times Eat Out Per Month			−0.0002
			(0.014)
			p = 0.990
Constant	3.759	3.109	3.875
	(0.100)	(0.142)	(0.506)
	p = 0.000***	p = 0.000***	p = 0.000***
Fixed Effects	No	No	Yes
Observations	243	243	242
R^2	0.046	0.171	0.231
Adjusted R^2	0.042	0.164	0.154
Residual Std. Error	1.073 (df = 241)	1.002 (df = 240)	1.006 (df = 219)
F Statistic	11.498*** (df = 1; 241)	24.804*** (df = 2; 240)	2.991*** (df = 22; 219)
Note:			*p<0.1; **p<0.05; ***p<0.01

Randomization Inference

The sharp null hypothesis states that the treatment effect is zero for every person, not just zero overall. In other words for each person, their potential outcomes to treatment and potential outcomes to control are exactly the same. Randomization inference assumes the sharp null hypothesis and allows us approximate the distribution of treatment estimates we would reach if the treatment had no effect.

We used randomization inference to simulate 100,000 different treatment assignments, and therefore 100,000 different treatment estimates. This yielded us with a p-value of *0.00097* meaning we would have less than a 0.1% chance of observing our treatment effect by random chance.

Chi-Squared Test

Chi-Square tests are used for categorical data to determine if there is a difference between various groups. Since our outcome measure is a Likert-scale rating, it is an ordered category and not a truly continuous variable. Therefore, the Chi-Square test is an appropriate test for our data.

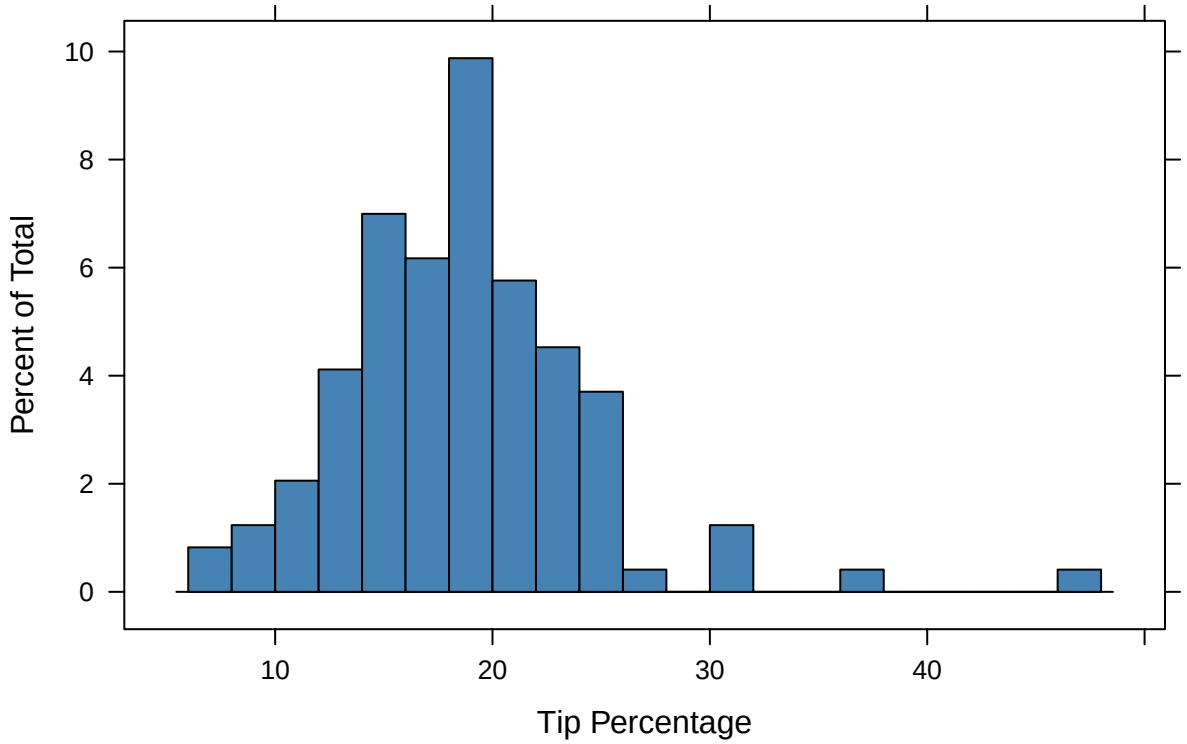
```
##
##  Pearson's Chi-squared test
##
## data:  table(ReasonablenessRating, Treatment)
## X-squared = 11.657, df = 4, p-value = 0.0201
```

The results of our Chi-squared test further confirm what we saw in the linear models and randomization inference. With a p-value of 0.02, we reject the null hypothesis that the responses from the treatment and control groups are independent. To say that another way, the treatment and control outcomes have a different distribution and therefore the treatment itself changed the potential outcomes from one distribution of satisfaction to a different distribution.

Overall Bill Price

One concern we had was that maybe people in the tipping group would decide to leave very small tips. With no ramifications for leaving a bad tip, they could decide to go that route and it would potentially influence our results. In Figure 2, we show the distribution of tip percentages left by respondents. There are a couple questionable values but overall the distribution of tip percentages looks good. The mean tip percentage is 18.7% (95% confidence interal of 17.7% - 19.8%) which is fairly normal therefore we feel that respondents treated the tips like they would in a normal restaurant setting.

Figure 2



Another concern of ours was that there would be different bill prices between the treatment and control group. In other words, the average bill prices for the treatment and control groups would be different because one group thought the prices were more reasonable than the other. The direction of this effect could go either way. Maybe one group feels the prices are so reasonable that they order more food, or they feel the prices are so unreasonable that they opt for the cheapest dish in the hopes of making the bill as tolerable as possible.

Figure 3

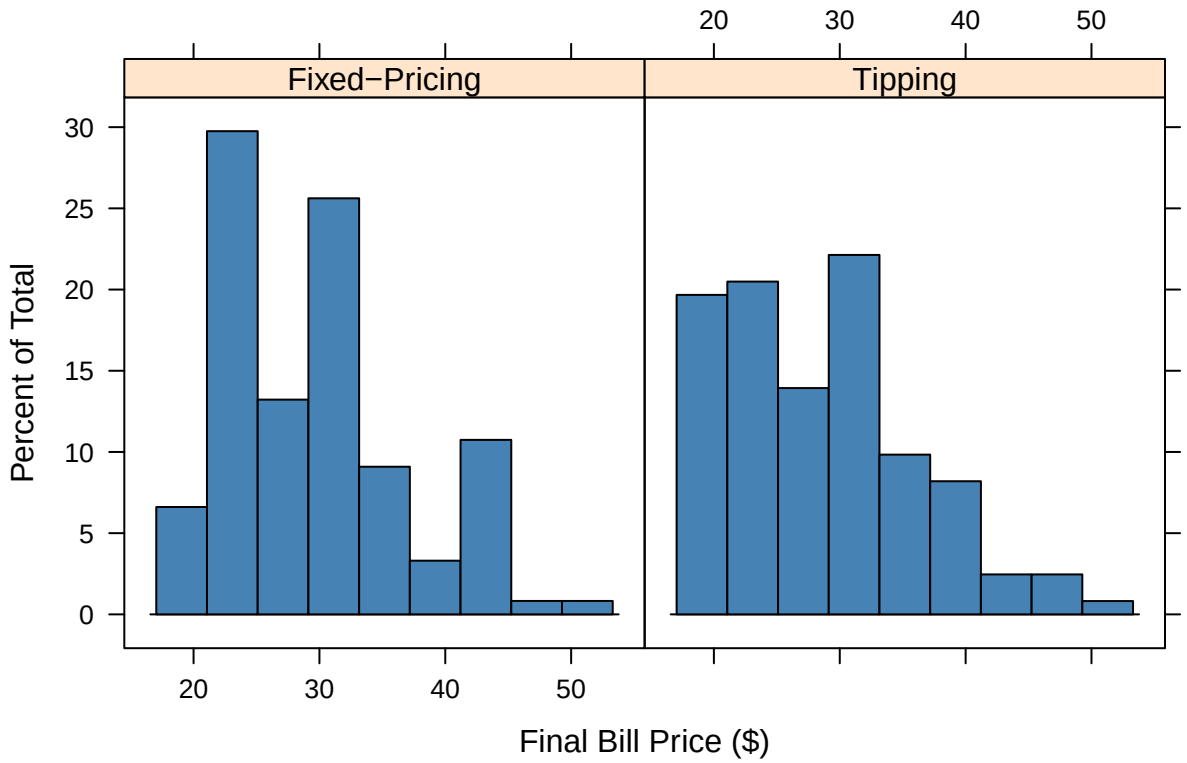


Figure 3 shows the total bill prices including meal, 7% tax, and tip (if applicable). The average total bill prices were $29.06 for the treatment group and $29.52 for the control group. This means that on average, the tipping group spent more money but was thought the cost was more reasonable. However, a t-test on the difference in bill prices between the two groups gives a p-value of 0.625249 meaning we cannot conclude that the bill prices were significantly different between the two groups.

Conclusion

Our initial hypothesis was that subjects would be less satisfied with all-inclusive pricing than standard tipping pricing, and we believe the data support that position. The regression model including all major covariates showed only two fields to be significant in predicting the subjects' views on meal price reasonableness: baseline meal spending and the treatment of all-inclusive pricing. The former seems consistent in that those higher spending may have had budgets beyond our meal pricing, and the latter supports our theory with the greatest coefficient in the model. We received further confirmation in the lack of significant difference in average meal price between treatment and contro, indicating that while the groups had similar spending habits, their customer satisfaction levels with pricing were indeed different.