

W241 Final Report

Vous voulez du vin? - Does the use of foreign language advertising in wine increase customer's purchase likelihood?

Rory Liu, Charlotte Swavola, Sharad Varadarajan

Introduction

When walking through the grocery store, one can easily spot products with foreign-language labels - "Vins de France" on a wine bottle, "Wirklich gut!" on a sausage pack, etc. Advertisers often confront consumers with foreign languages, such as German or French¹. In fact, the use of foreign language in advertising is a well-studied topic. According to the study done by France Leclerc, Bernd Schmitt and Laurette Dubé, showing French pronunciation of a brand name affects the perceived hedonism of the products, attitudes toward the brand, and attitudes toward the brand name². Furthermore, Jos Hornikx and Frank van meurs exclaim in their study that the use of foreign language serves as a strong cue for a product's country of origin, and the associations that the foreign language evoke and those that the country-of-origin evoke are similar³.

While scholars seem to agree that foreign language advertising shapes product perception by implicitly giving consumers cues on product's country of origin, few of the existing studies dive deep into the causal link between foreign language and consumer's purchase likelihood. This is an important causal link to establish. At the end of the day, the success of an advertising campaign is, or should be measured by the additional sales the campaign generates. Without an understanding of whether having foreign language in product communications increases sales, advertisers will be ill-guided in their decision to apply such advertising techniques.

In our study, we seek to understand the relationship between foreign language advertising and consumer willingness to pay. To narrow down our scope for a manageable experiment, given our time and resource constraints, we decided to focus on one particular product: wine. The high variance of wine prices, and the strong association between place of origin and the signal on quality makes it a great subject for our study.

Experimental Design

To fully understand the relationship between foreign language advertising and wine purchases, we drafted a multi-factorial design for our experiment. There are three main factors that we want to investigate: advertising language, wine's country of origin, and length of flavor profile description.

For advertising language, we will test 3 languages: English (as baseline), French and German. We selected French because France is typically perceived as a premium wine production country, and we

¹Jos Hornikx, Frank van Meurs & Robert-Jan Hof (2013) The Effectiveness of Foreign-Language Display in Advertising for Congruent versus Incongruent Products, *Journal of International Consumer Marketing*, 25:3, 152-165, DOI: 10.1080/08961530.2013.780451

²Leclerc, F., B. H. Schmitt, and L. Dubé. 1994. Foreign branding and its effects on product perceptions and attitudes. *Journal of Marketing Research* 31 (2):263-270

³Jos Hornikx & Frank van Meurs (2017) Foreign Languages in Advertising as Implicit Country-of-Origin Cues: Mechanism, Associations, and Effectiveness, *Journal of International Consumer Marketing*, 29:2, 60-73, DOI: 10.1080/08961530.2016.1243996

		Page Language					
		English		French		German	
		Flavor Profile Description					
		Long			Short		
Country of Origin	US	1	2	3	4	5	6
	France	7	8	9	10	11	12

Figure 1: Groups Design

hypothesize that advertising in French will increase consumer’s willingness to pay. We also selected German as a comparison language. Germany is not typically associated with wine, and therefore, by including German, we will be able to see whether the effect of having foreign language is limited to the language of the country associated with a product, or whether it expands to other languages as well.

For the wine’s country of origin, we included US and France. Here we are particularly interested in the interaction between country of origin and advertising language. Prior studies have found that foreign-language advertising serves as country of origin cues. By including country of origin as a variable, we are able to see whether foreign language has any additional impact when the country of origin is included.

For flavor profile description, we have two versions - long and short. We believe that this could be a proxy for foreign language “dosage”. A long flavor profile in French might be more noticeable than a short French tagline, and may amplify the effect that we find.

To summarize, we have a 3 x 2 x 2 design, which results in 12 treatment conditions. (See summary in figure 1).

Because of time and resource constraints, we could not conduct a real-life experiment with actual wines in actual stores. Instead, we conducted the study through a Qualtrics survey and gathered responses via Mechanical Turk. The survey is structured into the sections shown in Figure 2.

We start with some demographic question about participants’ primary language, income and gender. These will serve as covariates in our regression models. Then, we ask participants about their wine-related behavior, including how often they drink wine, whether they like Cabernet Sauvignon, and how often they purchase bottles of wine. We think that these wine-related behavior will explain some of the variation in willingness-to-pay, and therefore will be good covariates that help us reduce our standard error. The third section is the main part of our experiment where we introduce the treatment. Here, we randomly assign the participants into one of the twelve groups mentioned above, and show them a simulated purchasing page (example seen in Figure 3).

In the simulated wine page, we translate all texts into the groups assigned language, and provide english translation for key elements (e.g. flavor profile, ‘top wine of the year’ tag) to ensure that

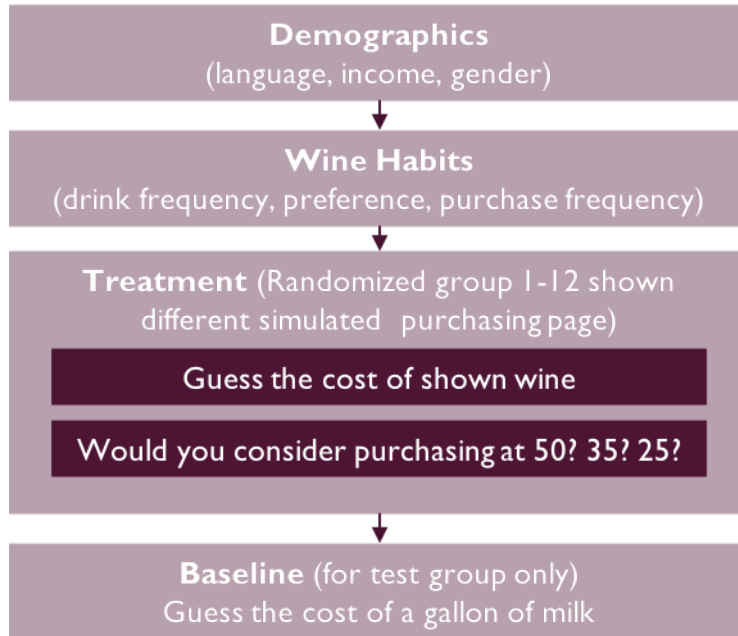


Figure 2: Survey Flow



Figure 3: Page Sample

the respondent understands the information. All groups are shown the same wine (which was a hypothetical wine fabricated by us), and the same descriptive information. We made the decision to present the page entirely in French or German for related tracks because we wanted to make sure participants noticed our language treatment.

We then ask the respondents to guess the cost of the wine. This is an un-anchored direct pricing question, aimed to understand respondents' first perception of the wine shown. Afterwards, we use a simplified version of the Gabor Granger method developed by André Gabor and C. W. J. Granger in the 1960s⁴. In this method, we ask respondents whether they would consider purchasing the shown wine at \$50. If respondents say no, we ask the same question at a lower price point of \$35. If they still answer no, we lower the price further to \$20. Through this series of 3 questions, we seek to understand where their true willingness-to-pay lies. We selected the 3 price points to cover a wide range of potential willingness-to-pay, and validated our choice through a pilot study of 100 respondents. From the pilot, we saw that most respondents' willingness to pay fell into the \$20-\$50 price range.

Our survey is concluded after the treatment section for our main sample of 1200 respondents. We used the data from this main sample to run various regressions and randomized inferences that sought to answer our research question.

However, to validate the findings from this main sample, we collected another (smaller) sample of 600 observations. The data from this second batch of respondents served as our validation test. By running the key models we deemed important (with statistically significant findings) from main sample with this validation data set, we can confirm that the relationship we found is real.

In our second launch, we maintained the same survey experience by asking the same questions in the same sequence. However, we did introduce a baseline question at the very end of the second survey, asking respondents to estimate the price of a gallon of milk. Because this question came at the very end of the survey, it should not interfere with our treatment, and would not create any difference between the data from first and second launch. However, having this baseline question helps us establish a reference price point for respondent's price guesses, and adjust for any overall inflated / deflated guesses.

EDA and Data Cleansing

After loading the data gathered from our survey, we first went through some general data processing and examined descriptive statistics to better understand our general results. Table 1 and Figures 4 & 5 below helped support the team in getting a high-level summary of the data. With regards to Figure 4, the team noticed considerable variance in respondent's price perception but that the large majority of guesses fell between \$1-\$100 dollars. This finding greatly influenced our model building decisions. With regards to Figure 5, the team was pleased to see a rather normal distribution for both purchase frequency and drink frequency. Additionally, we noticed that English was the primary language for almost all of our respondents, and that a menial number of respondents were fluent in either German or French; these are encouraging statistics considering that we want to evaluate the effect of languages that are "foreign" to our users.

⁴Gabor, A. and Granger, C. (1966). Price as an Indicator of Quality: Report on an Enquiry. *Economica*, 33(129), p.43.

Table 1: Output variables, Pre-Cleaning

Statistic	N	Mean	St. Dev.	Min	Max
Guess_the_Cost	1,207	75.243	339.562	1.000	8,600.000
Purchase_50	1,206	0.363	0.481	0	1
Purchase_35	1,206	0.532	0.499	0	1
Purchase_20	1,206	0.824	0.381	0	1

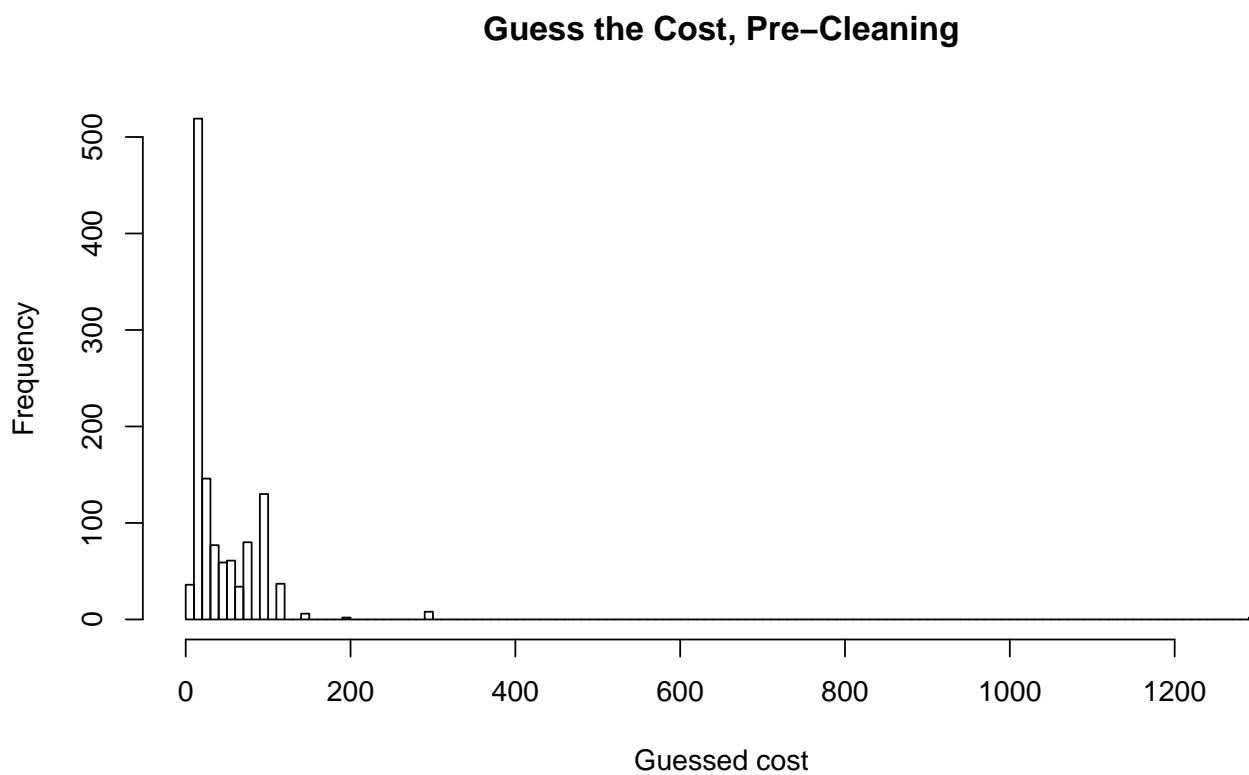


Figure 4: Guess the Cost, Pre-Cleaning

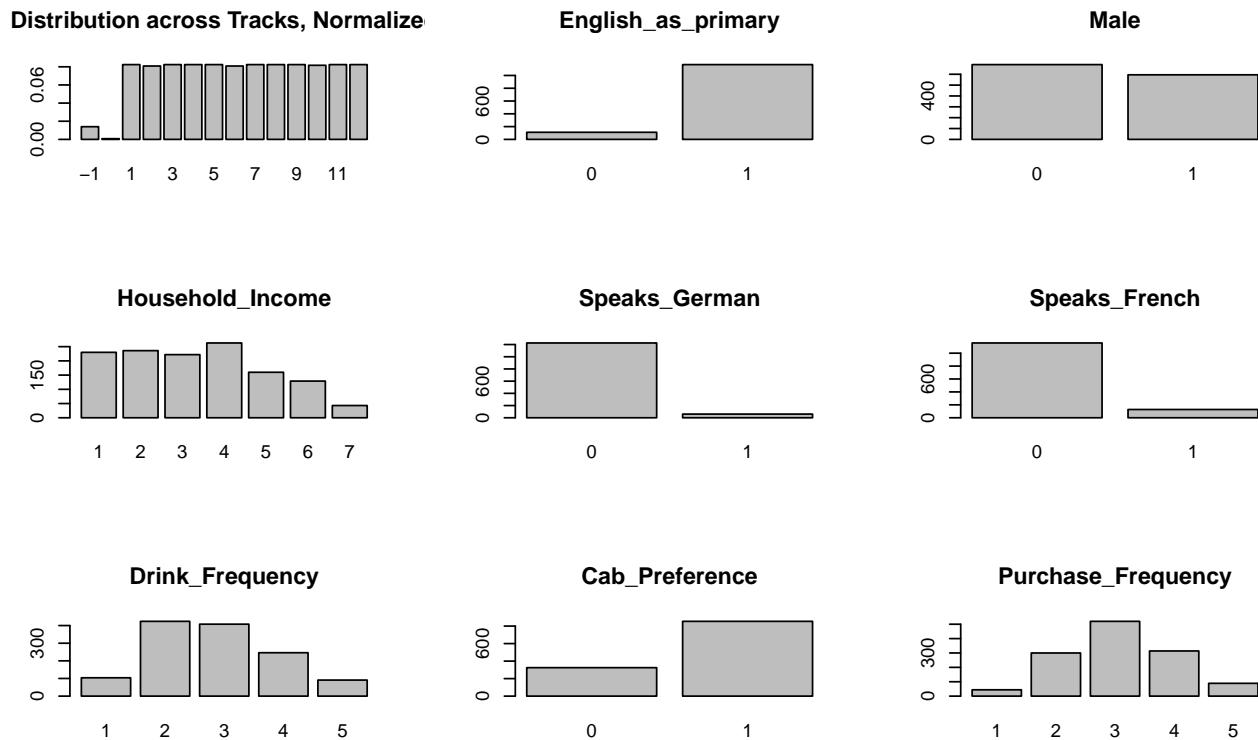


Figure 5: Univariate Analysis of Covariates- Pre-Cleaning

After evaluating these high-level statistics, we proceeded to analyze more granular details regarding each submission and discovered the following areas that require attention:

- 1). There was ~5% attrition (i.e. 5% of the respondents did not finish their survey). The attrition we see falls into 2 categories: pre-treatment attrition and post-treatment attrition. For pre-treatment attrition, we proceeded to clean them out completely, because these respondents were not exposed to treatment yet, and their dropping off can't have anything to do with treatment we introduce. Dropping them will not create bias in our data. For post-treatment attrition, we wanted to be more careful. We will examine these later on to understand whether we have differential attrition based on treatment assigned.
- 2). There was ~1/3 of respondents that have duplicated IP address and/or geographical coordinates (latitude and longitude). We are worried about these responses because they might be repeat survey takers (that potentially got exposure to more than one treatment), or they might be bots or click farms that do not answer surveys seriously. We proceed to clean out all responses that do not have a unique ip address and/or geographical coordinates.
- 3). As seen from Figure 4, we reiterate that we noticed many large outliers for guess-the-cost. The large variation in the answers will hurt our ability to obtain a reasonable standard error and find significant findings. We need to pay special attention in our modeling phase later, to try to minimize the effect of outliers.

After cleaning the data, we analyzed the previous descriptive statistics once more to see the effects. Below, in Table 2 and Figures 6 & 7, we see that cleaning reduced our outlier problem significantly for guess-the-cost, although considerable variation still exists.

Table 2: Output variables, Post-Cleaning

Statistic	N	Mean	St. Dev.	Min	Max
Guess_the_Cost	807	39.715	115.716	4.000	1,999.000
Purchase_50	806	0.261	0.439	0	1
Purchase_35	806	0.450	0.498	0	1
Purchase_20	806	0.793	0.406	0	1

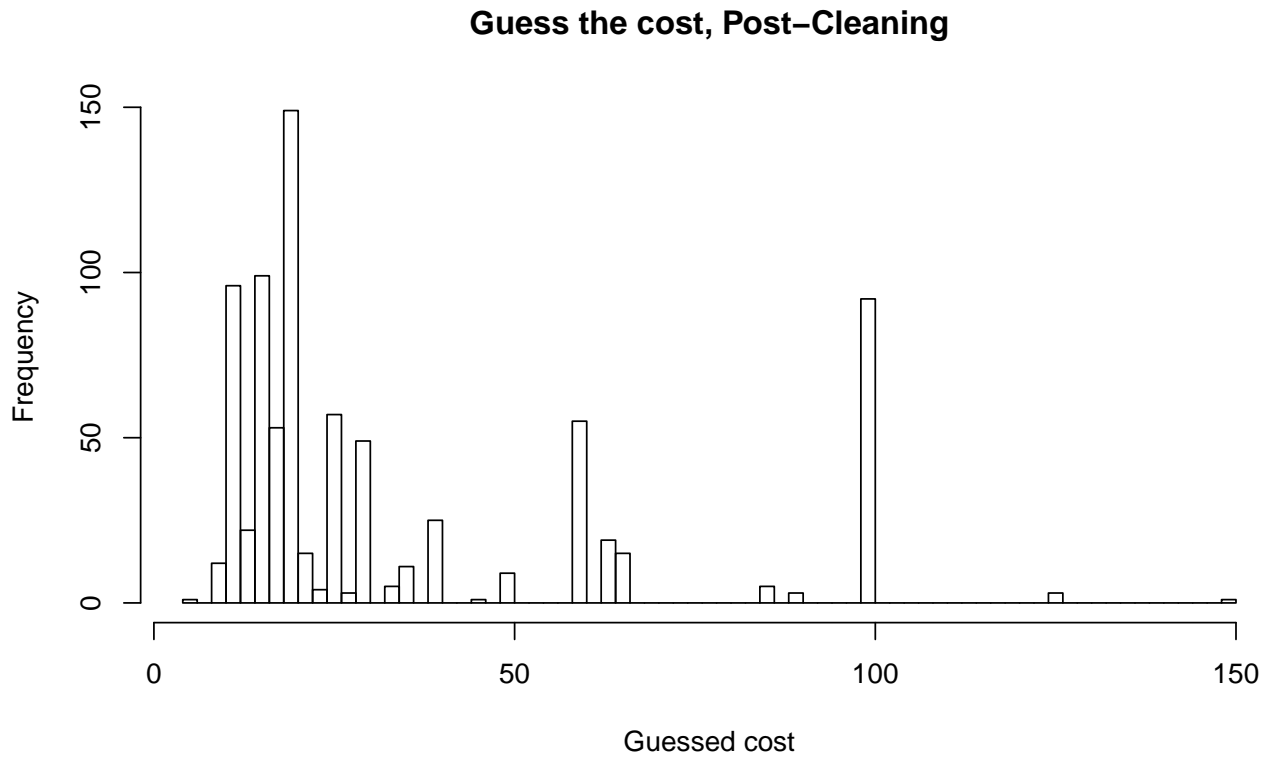
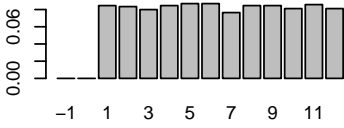
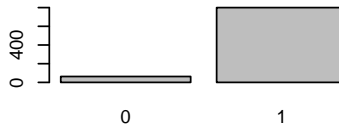


Figure 6: Guess the Cost, Post-Cleaning

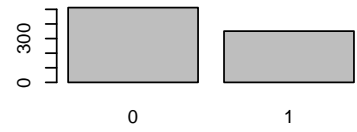
Distribution across Tracks, Normalize



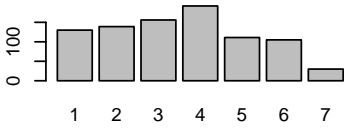
English_as_primary



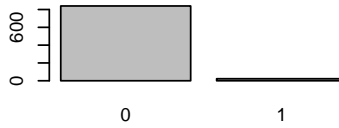
Male



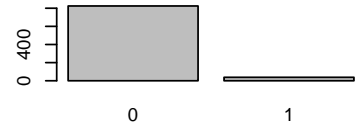
Household_Income



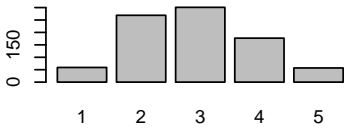
Speaks_German



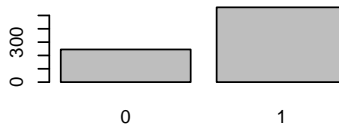
Speaks_French



Drink_Frequency



Cab_Preference



Purchase_Frequency

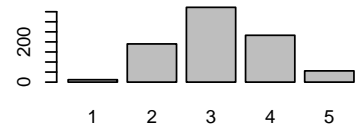
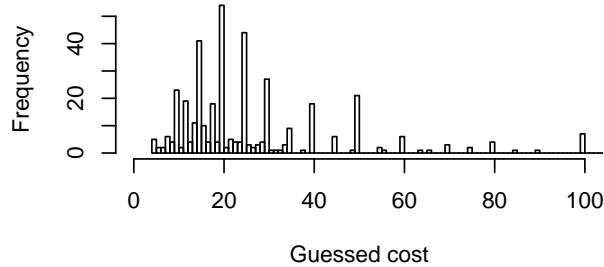


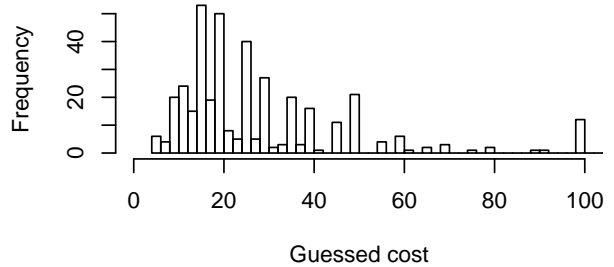
Figure 7: Univariate Analysis of Covariates- Post-Cleaning

Next, the team decided to dive a bit deeper into the guess-the-cost results, to see whether a treatment effect was obvious through histograms. We evaluated histograms, in Figure 8, that were summarized at each level of our three treatment factors:

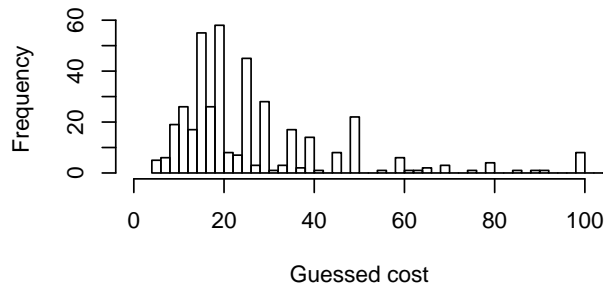
US Origin Wine



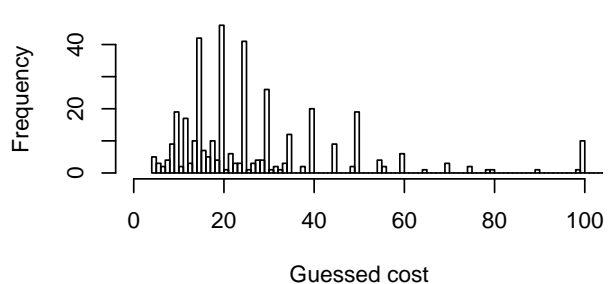
French Origin Wine



Short Description



Long Description



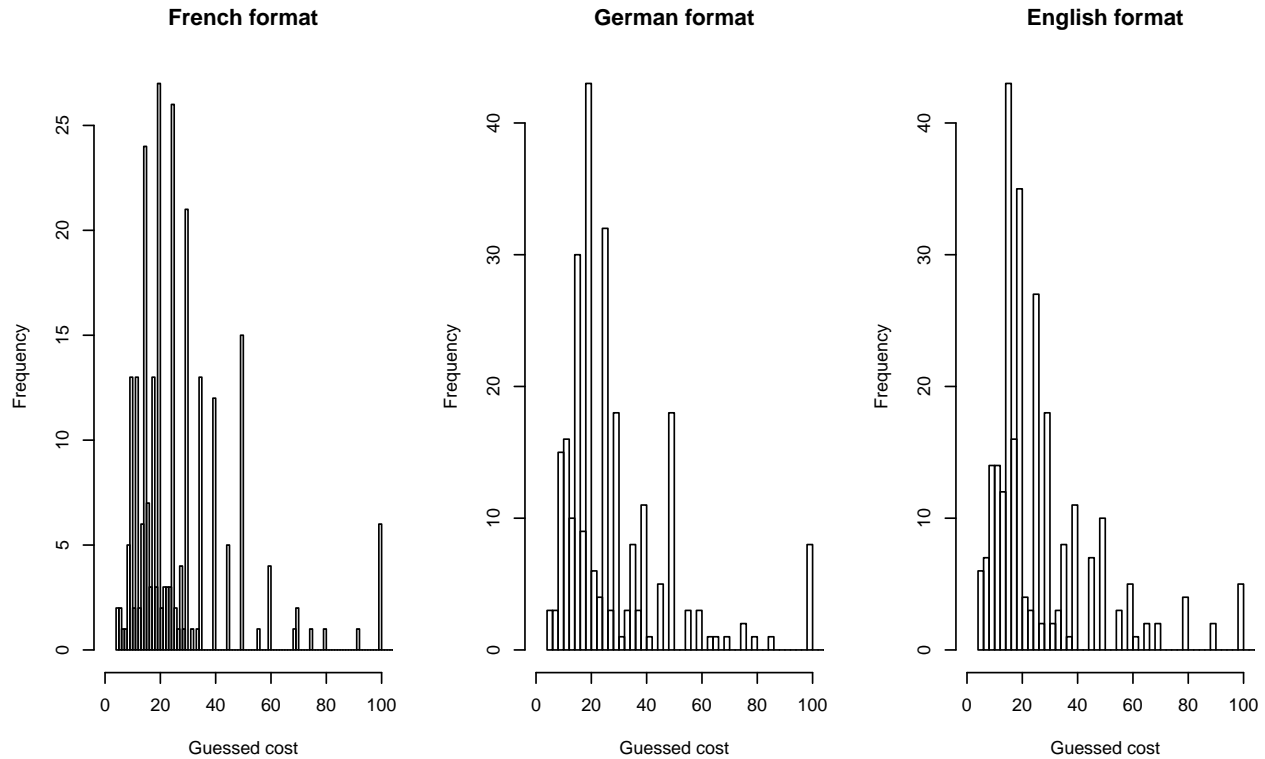


Figure 8: Guess The Cost- By Treatment

From these histograms of the answers, we don't observe a very apparent difference based on page language, country of origin, or length of description. This is not to say that we don't have any treatment effect, it's that the effect that our treatment has may be too subtle to appear in these visualizations.

Experimental Assumptions

Before beginning our modeling, the team proceeded to first validate our experimental assumptions, including conducting a covariate balance check, a check for potential differential attrition, and a check for potential heterogeneous treatment effects.

Demographic/Covariate Balance Check

For a covariate balance check, the team regressed each treatment factor against our covariates. The goal was to use the F-statistic as an indicator of whether our covariates had any explanatory power on which treatment track a respondent was assigned to. For 3 treatment factors (Origin, Language Format, Description length) with 2, 3 and 2 levels respectively, the team ended up constructing four regressions that are listed below:

$$1) FrenchPage = \beta_0 + \beta_1 Male + \beta_2 Income + \beta_3 EnglishPrimary + \beta_4 DrinkFrequency + \beta_5 CabPreference + \beta_6 PurchaseFrequency + u$$

$$2) \text{GermanPage} = \beta_0 + \beta_1 \text{Male} + \beta_2 \text{Income} + \beta_3 \text{EnglishPrimary} + \beta_4 \text{DrinkFrequency} + \beta_5 \text{CabPreference} + \beta_6 \text{PurchaseFrequency} + u$$

$$3) \text{USOrigin} = \beta_0 + \beta_1 \text{Male} + \beta_2 \text{Income} + \beta_3 \text{EnglishPrimary} + \beta_4 \text{DrinkFrequency} + \beta_5 \text{CabPreference} + \beta_6 \text{PurchaseFrequency} + u$$

$$4) \text{LongDescription} = \beta_0 + \beta_1 \text{Male} + \beta_2 \text{Income} + \beta_3 \text{EnglishPrimary} + \beta_4 \text{DrinkFrequency} + \beta_5 \text{CabPreference} + \beta_6 \text{PurchaseFrequency} + u$$

In these regressions, Male, EnglishPrimary and CabPreference are binary variables. Income, Drink Frequency and Purchase Frequency represent factor vectors. Equations 1 and 2 were evaluated on subsets of our original data frame. For equation 1, the data frame only contained observations where the purchasing page was in English or French. For equation 2, the data frame only contained observations where the purchasing page was in English or German. After running all four of these regressions, the p-values for each of the F-statistics were much greater than 0.05. The team concludes that our covariate balance check passes, seeing that we cannot reject the null hypothesis that all beta values are zero for these regressions.

Attrition check

As mentioned earlier, our attrition is rather low for the entire study (~5%), so we are not very concerned about potential bias caused by possible differential attrition. However, we decided to do a check for completeness. Here, we used logistic regression (since whether or not a respondent finishes the survey is binary) to check whether being assigned to certain treatment leads to more or less attrition than others.

Table 3: Attrition Regressions-Logistic with Odds Ratio

	<i>Dependent variable:</i>		
	Finished		
	(1)	(2)	(3)
French_Purchasing_Page1	0.326*** (0.396)		
German_Purchasing_Page1	0.405** (0.405)		
US_Origin1		0.990 (0.274)	
Long_Description1			0.653 (0.279)
Constant	30.556*** (0.339)	14.214*** (0.196)	17.826*** (0.214)
Observations	863	863	863
Log Likelihood	-205.143	-209.964	-208.780
Akaike Inf. Crit.	416.287	423.928	421.560

Note: *p<0.1; **p<0.05; ***p<0.01

After exponentiating the output coefficients, we saw from Table 3 that respondents assigned the French or German purchasing page were around 3 times less likely to finish the survey (compared to those assigned the English purchasing page). In other words, we do find evidence for differential attrition (this statistically significant finding was replicated with linear regression as well). However,

since our attrition rate is very low, even if the 5% that dropped out have very extreme outcomes, they are unlikely to bias our estimate too much. Therefore, we are not very concerned about this effect. Additionally, when conducting a differential attrition check for our validation study, the team was not able to replicate the finding, and found no statistically significant effects.

Heterogeneous treatment effects

We tested for heterogeneous treatment effects using regression for those who answered “Never” for their wine purchasing behavior. We hypothesized that people that never purchase wine may have a rather un-firm value perception of wine, and may be influenced by our treatment more easily. In Table 4, we observe a significance in the interaction term between never-purchase and German language, which does suggest the presence of HTE.

Table 4: HTE

	<i>Dependent variable:</i>		
		Guess_the_Cost	
	(1)	(2)	(3)
French_Purchasing_Page	0.842 (1.290)		
German_Purchasing_Page	1.301 (1.275)		
US_Origin		-1.262 (1.057)	
Long_Description			1.055 (1.072)
Never_purchased	9.218** (3.730)	13.656*** (3.070)	14.007*** (3.240)
French_Purchasing_Page:Never_purchased	-0.573 (5.361)		
German_Purchasing_Page:Never_purchased	12.145** (5.357)		
US_Origin:Never_purchased		-3.234 (4.431)	
Long_Description:Never_purchased			-3.675 (4.493)
Constant	23.786*** (0.900)	25.209*** (0.753)	24.078*** (0.751)
Observations	807	807	807
Residual Std. Error	14.281 (df = 801)	14.516 (df = 803)	14.630 (df = 803)

Note:

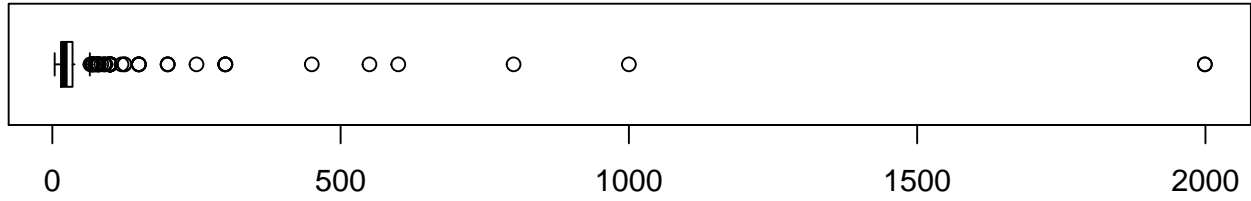
*p<0.1; **p<0.05; ***p<0.01

Modeling

RI with limit on guesses

Our first attempt to isolate a treatment effect was by using randomization inference and the sharp null hypothesis. By nature of RI, in which a high outlier may swing the entire treatment or control group, we limited the guesses to the bottom 95%. This does not bias our results because in RI, the eliminated outliers have equal chances of falling into treatment or control. Using the reduced outcome set, we tested the simple treatments and our specified interaction (same format*origin) hypotheses, the results of which can be seen in Figures 9 and 10.

Boxplot for Guesses, including outliers



Boxplot for Guesses, bottom 95 percent of guesses

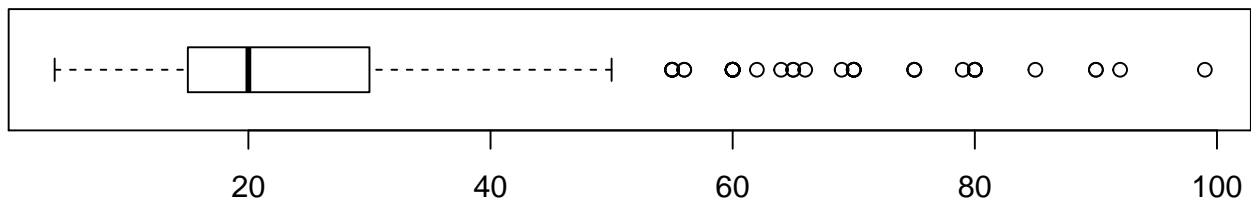


Figure 9: RI Boxplots

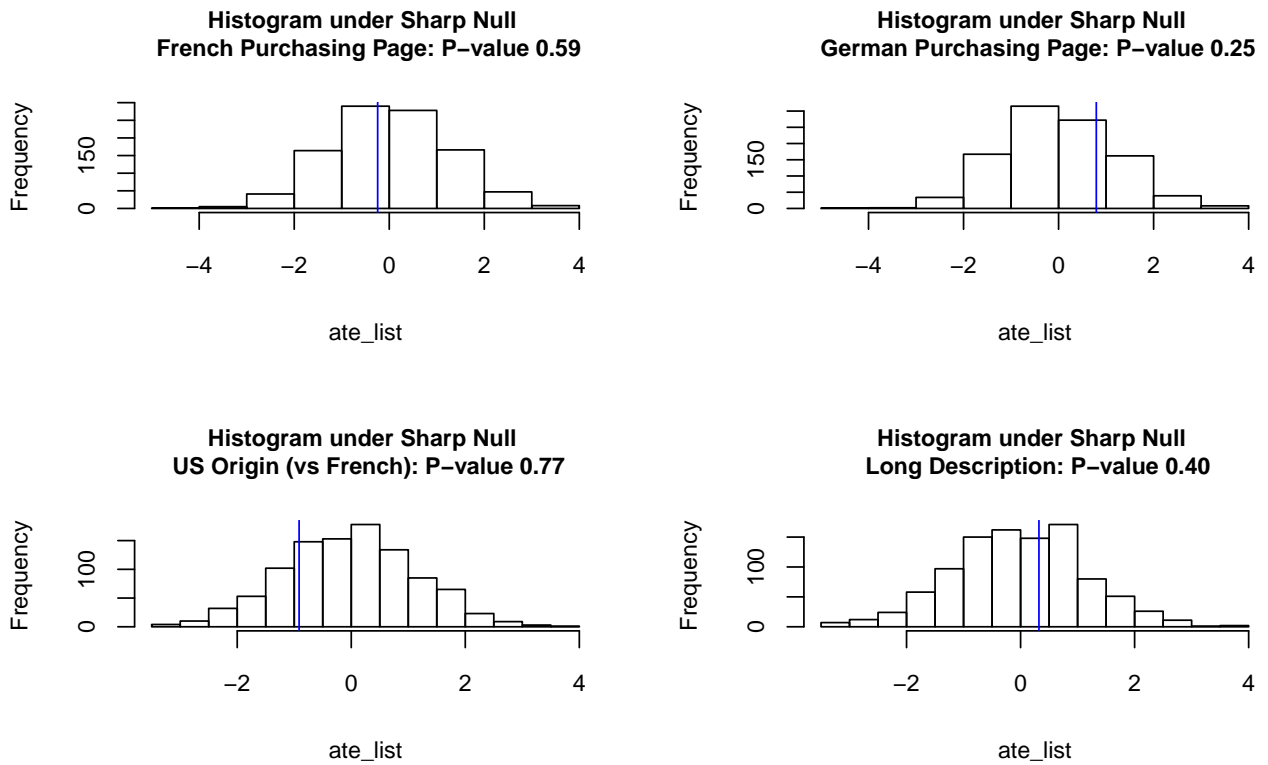


Figure 10: RI Sharp Null Hypothesis

The results from our Randomization Inference attempt are displayed in Figure 10. Here we see no statistically significant results for all of the 4 treatments (French purchasing page, German purchasing page, US origin, and Long description). Therefore, we could not reject the sharp null hypotheses that the treatment effect is zero for all respondents. The team moved on to regression models.

2 Model Evaluation Strategy

The first outcome of interest was respondents guesses for the price of a wine. Initially, the team wanted to see how standard OLS regressions would perform with our data. We were not very optimistic based on the considerable variance observed from our earlier descriptive statistics. Preliminary models that we built reinforced this concern. Table 5 contains simple OLS models where we used only the treatment variables as predictors. As you can see from the table, the most simplistic of regressions were returning large standard errors.

Table 5: Standard OLS Regressions

	<i>Dependent variable:</i>			
	Guess_the_Cost			
	(1)	(2)	(3)	(4)
French_Purchasing_Page1	-11.295 (10.030)			-11.426 (10.034)
German_Purchasing_Page1	-3.094 (9.905)			-2.892 (9.910)
US_Origin1		-8.508 (8.157)		-8.659 (8.168)
Long_Description1			-5.034 (8.161)	-4.877 (8.165)
Constant	44.413*** (6.985)	44.046*** (5.804)	42.213*** (5.720)	51.167*** (9.015)
Observations	806	806	806	806
R ²	0.002	0.001	0.0005	0.004
Adjusted R ²	-0.001	0.0001	-0.001	-0.001
Residual Std. Error	115.833 (df = 803)	115.780 (df = 804)	115.831 (df = 804)	115.869 (df = 801)
F Statistic	0.672 (df = 2; 803)	1.088 (df = 1; 804)	0.381 (df = 1; 804)	0.710 (df = 4; 801)

Note: *p<0.1; **p<0.05; ***p<0.01

After running a few more models with covariates, the team realized that these large standard errors were going to be an issue in all OLS regression models we run. Therefore, the team decided to go with a different strategy to evaluate a treatment effect for this outcome, and that was through Robust Linear Models. The idea of this model is to weigh observations differently based on how high their residuals are. We viewed this a compromise between eliminating extreme guesses entirely from the analysis and treating all price guesses equally like we would in OLS regression. When re-running the same models with RLM, where treatment variables were the only predictors, we saw that the treatment effect dropped considerably (in Table 6); however the standard errors were more reasonable and reflective of the distribution among the majority of observations.

For the second outcome of interest, the team decided to evaluate three different outcomes of willingness to purchase. Since our outcomes for each of these 3 price points are binary, we decided to use binomial logistic regression as our method of analysis. With regards to the output of logistic regression models, the team chose to exponentiate the coefficients, as we could then interpret effects as the odds of purchasing the wine at a certain price point (odds ratio). The team also considered using an ordinal logistic regression instead, and looking at maximum pricepoint at which individuals

Table 6: RLM Basic Regressions

	<i>Dependent variable:</i>			
	Guess_the_Cost			
	(1)	(2)	(3)	(4)
French_Purchasing_Page1	0.837 (1.274)			0.842 (1.248)
German_Purchasing_Page1	1.611 (1.258)			1.640 (1.232)
US_Origin1		-1.478 (1.023)		-1.561 (1.016)
Long_Description1			1.009 (1.031)	1.066 (1.015)
Constant	24.300*** (0.887)	25.656*** (0.728)	24.480*** (0.723)	24.429*** (1.121)
Observations	806	806	806	806
Residual Std. Error	14.634 (df = 803)	13.840 (df = 804)	14.103 (df = 804)	14.084 (df = 801)

Note:

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

		Model type & treatment variable							
		Simplistic Model (1 dimension only)				Isolated Interaction	Interactions		
		German Format	French Format	US Origin	Long Description	Same Origin* Format	Pairwise interaction model	Fully Saturated (All interactions)	
Outcome Variable	\$ Guess	No statistically significant findings					* Long Description (+) ^ Long*US Origin (-) ^ German Format (+)	^ Long Description (+)	
	\$ Guess (With CV)						^ Long Description (+) ^ Long*US Origin (-)		
	\$50 Purchase Likelihood (With CV)								^ German Format * US Origin (-) ^ German Format * US Origin * Long (+)
	\$35 Purchase Likelihood (With CV)								2 * German Format * Long (-) * German Format * US origin * Long (+) ^ German format (+)
	\$20 Purchase Likelihood (With CV)						1 * French Format * French Origin (+) ^ Same Format & Origin (+)	* French Format (*) ^ French Format * US Origin (-)	

Legend ^10% statistical significance
 * 5% statistical significance

Figure 11: Model summary

would purchase; however we did not believe the levels of our willingness to purchase outcome met the conditions of the proportional odds assumptions. Therefore we thought it would be more prudent to use binary logistic regression with three separate outcomes.

Model Matrix

Figure 11 describes the different RLM and Logistic regression models we ran on our training set. Each column describes the details of predictors we used in the model, and each row describes the outcome variable for the model and whether or not covariates were included:

- 1) The left-hand side of the table pertains to our most simple models, where we only used the treatment variables as predictors without any interactions
- 2) For Isolated Interactions we only examined the effect of treatment where the wine origin and purchasing page format are the same. For example the effect of French origin-French language

treatment, or US origin-English language treatment.

- 3) For pairwise interaction models, we would include only 2 of 3 treatments per model and observe their dynamic. For example, one model could include language format and description length as predictors, as well as their interaction, regardless of the origin of the wine.
- 4) For the fully saturated models, we included each of the treatment variables, and all pairwise and three-prong interactions between treatment.

On the left-hand side of Figure 11 there is a box stating, “No Statistically Significant Findings” and that pertains to the most simplistic regressions we ran. For example, when trying to isolate the effect of wines of US Origin, regardless of the length of the flavor profile or purchasing page format, we saw no significance. As we move left to right on Figure 11 we come across some of our more detailed models. At this point, the team started to uncover some findings that are significant at the 10% level and others at the 5% level. For brevity, we will discuss the two findings we found most interesting. The first was the positive effect of the interaction between French Format and French Origin on the willingness to purchase at \$20.

$$Purchase_{20} = \beta_0 + \beta_1 FrenchFormatOrigin + \beta_2 EnglishFormatOrigin + \beta_3 Male + \beta_4 Income + \beta_5 EnglishPrimary + \beta_6 DrinkFrequency + \beta_7 CabPreference + \beta_8 PurchaseFrequency + u$$

The results in Table 7 indicate that a respondent was 2.075 times more likely (statistically significant at 5%) to purchase a wine of French Origin-French Format for \$20 than a respondent who saw a wine of English Origin-French Format. This made a lot of sense to us, considering that adding French language to describe a French wine could possibly attribute more authenticity to the wine.

The second model of interest was the fully saturated model where there was a negative effect involving German Format on the willingness to purchase at \$35.

$$Purchase_{35} = \beta_0 + \beta_1 FrenchPurchasingPage + \beta_2 GermanPurchasingPage + \beta_3 USOrigin + \beta_4 LongDescription + \beta_5 FrenchPurchasingPage * USOrigin + \beta_6 GermanPurchasingPage * USOrigin + \beta_7 FrenchPurchasingPage * LongDescription + \beta_8 GermanPurchasingPage * LongDescription + \beta_9 USOrigin * LongDescription + \beta_{10} FrenchPurchasingPage * USOrigin * LongDescription + \beta_{11} GermanPurchasingPage * USOrigin * LongDescription + \beta_{12} Male + \beta_{13} Income + \beta_{14} EnglishPrimary + \beta_{15} DrinkFrequency + \beta_{16} CabPreference + \beta_{17} PurchaseFrequency + u$$

The results in Table 8 indicate that a respondent was about 4 times less likely (statistically significant at 5%) to purchase a wine with German Language-Long Description-French Origin for \$35 than a respondent who saw a wine of English Language-Long Description-French Origin. Since German is not a language typically associated with wine, the team was curious whether German was the driving factor in respondents choice not to purchase the wine at \$35.

Validation Study

Based on the size of the matrix in Figure 11, it was evident that our team evaluated many regression models. Therefore, making any causal inferences from these model results would not be prudent,

Table 7: Log Regression- Same Format/Origin

	<i>Dependent variable:</i>
	Purchase_20
French_format_wine	2.075** (0.311)
English_format_wine	1.136 (0.265)
Male1	1.485* (0.209)
Household_Income2	1.381 (0.349)
Household_Income3	1.254 (0.340)
Household_Income4	1.150 (0.321)
Household_Income5	1.260 (0.368)
Household_Income6	1.948 (0.425)
Household_Income7	0.611 (0.523)
English_as_primary1	1.989* (0.355)
Drink_Frequency2	1.591 (0.437)
Drink_Frequency3	2.041 (0.500)
Drink_Frequency4	0.989 (0.558)
Cab_Preference1	1.650** (0.208)
Purchase_Frequency2	0.0000 (436.329)
Purchase_Frequency3	0.0000 (436.329)
Purchase_Frequency4	0.0000 (436.329)
Purchase_Frequency5	0.0000 (436.330)
Constant	838,162.600 (436.329)
Observations	758
Log Likelihood	-335.509
Akaike Inf. Crit.	709.018

Note: *p<0.1; **p<0.05; ***p<0.01

Table 8: Log Regression Fully Saturated

	<i>Dependent variable:</i>
	Purchase_35
French_Purchasing_Page1	1.293 (0.378)
German_Purchasing_Page1	1.866* (0.368)
US_Origin1	1.317 (0.364)
Long_Description1	1.854 (0.378)
Male1	1.646*** (0.157)
Household_Income2	0.959 (0.284)
Household_Income3	0.688 (0.281)
Household_Income4	0.711 (0.270)
Household_Income5	0.861 (0.299)
Household_Income6	0.828 (0.306)
Household_Income7	0.548 (0.499)
English_as_primary1	1.829* (0.325)
Drink_Frequency2	0.931 (0.346)
Drink_Frequency3	0.911 (0.391)
Drink_Frequency4	0.608 (0.455)
Cab_Preference1	1.717*** (0.175)
Purchase_Frequency2	0.283 (0.861)
Purchase_Frequency3	0.359 (0.879)
Purchase_Frequency4	0.455 (0.911)
Purchase_Frequency5	0.169 (1.234)
French_Purchasing_Page1:US_Origin1	1.080 (0.527)
German_Purchasing_Page1:US_Origin1	0.479 (0.517)
French_Purchasing_Page1:Long_Description1	0.724 (0.538)
German_Purchasing_Page1:Long_Description1	0.272** (0.538)
US_Origin1:Long_Description1	0.630 (0.524)
French_Purchasing_Page1:US_Origin1:Long_Description1	0.942 (0.748)
German_Purchasing_Page1:US_Origin1:Long_Description1	4.621** (0.741)
Constant	0.743 (0.936)
Observations	758
Log Likelihood	-499.050
Akaike Inf. Crit.	1,054.101

Note:

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

considering the chances of retrieving a false positive (type II error) increases considerably when running all of these models. To avoid making false causal claims and verify that our findings were legitimate, the team re-evaluated our two findings of interest on a validation set, to see if we could replicate the direction of the effect with statistical significance.

As mentioned in the introduction, the team added a baseline price-perception question for a gallon of milk post-treatment in the validation study. By adding this baseline question, the team was able to gauge whether or not a respondent was providing guesses for cost in terms of US currency; we chose to remove all observations where respondents guessed over \$10 for a gallon of milk, with the hope of ridding ourselves of harmful outliers for our outcome variables. While this method helped remove many inflated guesses, there were still a number of outliers remaining in our dataset. Therefore, the team continued with RLM instead of OLS regression for our Guess the Price outcome variable.

After data collection and data cleansing for our validation set, the team went through the following tasks prior to model analysis:

- 1) Conducted a covariate balance check and confirmed there were no baseline differences between the groups (in terms of our measured covariates)
- 2) Observed that the differential attrition identified in our training set did not exist in our validation set. In fact, respondents who were assigned the French/German purchasing page treatments were less likely to attrit in our validation study compared to those assigned the English purchasing page (though not statistically significant). The direction of this differential attrition effect is opposite to what was observed in our training study.
- 3) The heterogeneous treatment effect for never-purchasers did not persist in the validation study as well

After re-running the two models of interest from the training study, we observed the following results on our validation set:

- 1) Respondents were 2.252 times more likely (statistically significant at 10%) to purchase a wine of French Origin-French Format for \$20 than respondents who saw a wine of English Origin-French Format. While the direction and relative magnitude of the effect persisted from the main study, the statistical significance dropped from 5% to 10% in the validation study.
- 2) Respondents were 1.592 times more likely (not statistically significant) to purchase a wine with German Language-Long Description-French Origin for \$35 than respondents who saw a wine of English Language-Long Description-French Origin. Compared to what we observed in the training study, the direction of the effect flipped and the magnitude of the effect dropped considerably. Additionally, the finding was no longer statistically significant.

Conclusion

The one significant finding that we have that persisted in both training and validation study (although only at 10% in the validation study) is the positive impact of French wine and French purchasing page. We found that this combination drives up purchase likelihood at \$20. Advertising in French for an entry-level French wine may be a good idea for businesses, though we recommend further validation.

Having gone through the study, we learned a lot about experimental design, and if we were to do the study again, we would improve our design in the following ways:

1. Introducing a baseline question in the training study. The milk-cost baseline question helped us control for some outlier observations, especially those caused by currency perception. Having this question in the training study would definitely help in our modeling process.
2. Limiting the geographical location of respondents. In mechanical Turk, we did not limit to only US respondents, for fear that we may not achieve our ambitious sample on time. If time permits, this limitation would be useful because it will then eliminate currency bias.
3. Increasing sample size, especially in the validation study. The large variation we see in responses also calls for large sample sizes. If we have enough time and resources, gathering a larger sample helps increase statistical power.
4. Changing treatment dosage. For this treatment, we used the “maximum dosage possible” by translating everything into the treatment language. This may have overwhelmed or confused respondents, which lead to some differential attrition in our training study. In a future study, it may be wise to reduce the treatment dosage in a more manageable way (e.g. only translating flavor profile and taglines),
5. Redesigning the guess-the-cost question. We liked the question because it is completely unanchored, which allowed us to find psychological price thresholds (like \$20). However, this type of question is outlier prone. In a new study, we may consider changing this question type, or at least putting a maximum answer threshold on guess-the-cost questions implicitly, without anchoring respondents. Having an implicit max price (the highest price imaginable for a wine, say \$300), where only people that entered more than that max value would see a validation, may help us reduce the magnitude of outliers.