The Effect Of First-Person Language On Responses To Online Classified Ads

Kyle Hamilton, Jonathan Landesman, Rajagopalan Mahadevan,
Daniel Sheinin, Umber Singh

w241 – University of California, Berkeley

December 20, 2015

TABLE OF CONTENTS

Abstract

We designed a field experiment to study the effects of using first person language in an online ad on the number of responses to the ad. Secondary outcomes include number of offers as well as offer amount. We present a unique way of randomizing online ads that avoids spillover effects. The mean response count, mean offer count, and the mean of the mean offer amount were all higher for the treatment group over the control group. Detailed R analysis is presented. Machine learning was also used to probe the connection between treatment and outcomes. There were no statistically significant results seen and so we did not see any significant effect from the use of first person language in an online ad.

Keywords: Online ads, Randomization, R, Machine Learning

The Effect Of First-Person Language On Responses To Online Classified Ads

This experiment gauges the extent to which first-person language in the text of a classified ad has an impact on the number of responses to the ad. Our experiment was inspired by a prior study by Doleac and Stein[1] on the effect of race on market outcomes by selling iPods through local online classified advertisements. We initially set out to measure the effect of gender on average offer in online classified ads on Craigslist, among the largest classified ad platforms in the United States.

In an **ideal experiment** we would have posted all ads in all locations on the same day, varying only whether the product in the photograph was held by a female hand versus a male hand. Upon conducting several pilot studies, it became apparent that the ideal experiment was not feasible for a number of reasons outside of our control. First, posting identical ads on Craigslist in multiple locations goes against the Craigslist terms of use. Not only is this prohibited, it’s technically almost impossible given the Craigslist mechanisms of detecting and blocking such attempts. Second, third party services claiming to get around these issues were prohibitively expensive, and even then, would not have been able to meet all of the requirements of the experimental design. Thus we revised our design to work within the Craigslist constraints as described in the experimental design section.

Ultimately, we attempted to estimate the effect that using first-person language in Craigslist ads would have on customer response. We used matched pairs of ads that differed only

[1] Doleac, J. L., & Stein, L. C. (2013). The visible hand: Race and online market outcomes. *The Economic Journal*, *123*(572), F469-F492.

in their use of language, posted them across multiple cities and measured the response. Our test was two-tailed: we hypothesized that the response would be different from zero, but we did not specify the direction of the effect.

I. Experimental Design

Working within the limitations of Craigslist, we created 50 different ad pairs, to be posted in locations matching each poster's geographic location. Each pair consists of a control ad, and a treatment ad. Since our group is made up five individuals, we were able to leverage five Craigslist locations, namely the San Francisco Bay Area, New York City, Los Angeles, Seattle, and Central New Jersey.

RANDOM ASSIGNMENT PROCEDURE

The "subjects" in this experiment were the combinations of ads and their posting contexts (city and day). Treatment was assigned by randomly allocating the 50 control and 50 treatment ads to posting contexts.

Complete randomization was implemented using R as follows. First treatment ads, then control ads were randomly distributed evenly across posting cities. In order to avoid spillover, a check was performed to ensure that the treatment and control ads from the same pair were not assigned to the same city. If the check failed, the control ad random assignment was repeated. (This method is feasible for 100 samples but would not scale well to larger sample sizes.) After city assignment, the ads for each city were randomly distributed evenly across posting days. The resulting distribution by city and day can be seen in Tables 1 and 2, respectively.

This is effectively a matched pairs design. For each ad/context combination, the non-text features of the ad itself represent the features that are matched between the two members of each pair. The posting context represents unmatched characteristics. In a typical matched pairs implementation, the subject, here the ad and posting context together, would have been randomly assigned to the treatment or control group as a whole. Instead, we randomly assign the non-matched characteristics (the posting context) to the ads, which had been pre-assigned to treatment or control groups.

TREATMENT DELIVERY

We each wrote twenty ads, ten of control, and ten of treatment. Control ads were written in third-person, while treatment ads were written to be as similar as possible to the control ad while using the more personal first-person form. The same product image was used in each ad pair. To avoid automatic removal by Craigslist for posting too many ads at the same time, we staggered our postings over 4 periods consisting of 5 ads each, and lasting 48 hours each. **Manipulation checks** were conducted by ensuring that the posted ads were visible in Craigslist search results - all posted ads in the final design were visible in the search results for each corresponding location.

OUTCOME MEASUREMENT

The primary outcome was the number of email responses to each ad. Given this straightforward measurement, no possibility of **measurement asymmetry** exists. Note that a measure of 0 does **not** indicate missingness; it simply indicates that the ad received zero responses. If an ad received multiple emails from the same bidder, this was measured as a single response. Issues of non-compliance and attrition are not relevant to the experiment, as cities and days cannot choose to comply nor attrite.

Secondary outcomes were the number of responses that contained offers and the average offer amount for each ad. With our sample size, and since the number of offers is by definition not greater than (and almost certainly less than) the number of responses, we were not optimistic about observing statistically significant results with our secondary outcomes, but it was important to include them as indicators of the responses' intent. For example, if the number of responses turned out to be higher in the treatment group than in the control group, this could be attributed to more interest in acquiring the product, or alternatively to the sense that a good deal would be easier to come by, either of which would be plausibly associated with the first-person wording. A correspondingly higher average offer for treatment ads would support the former interpretation while a lower average offer would support the latter.

II. Assumptions

Due to the craigslist restrictions described above, our experiment was forced to assume the following:

(i) Craigslist shopping populations are the identical (in expectation) across days; therefore a treatment ad posted on day 1 of the experiment will be seen by the same population with the same propensity to respond as a control ad posted on day 2 of the experiment.

(ii) Thanksgiving/ black friday had no effect on the likelihood of receiving a response. Due to timing and logistical constraints, the first day of our experiment was the Wednesday before Thanksgiving, and the second was the Friday afterwards.

(iii) The treatment generates an identical response across products and across product categories. We assume, in other words, that an iPod advertisement written in the first person and an vaccum cleaner ad written in the first person have identical effects on the viewer.

(iv) Ad copy as written by various authors represent equivalent implementation of the treatment.

These assumptions allow us to meet the criteria of excludability and non-interference. Presuming that there are no effects arising from the calendar day (assumption 2 above), our experiment is excludable, i.e. there is no reason to believe anything other than the treatment affected our results. Similarly, the deliberate geographic spread (randomization across cities) avoided spillover - it is unlikely that an individual viewed both treatment and control ads. It is unlikely that a craigslist shopper viewing our treatment ads in one city impacted another craigslist shopper viewing our control ads in a completely different city.

III. Data and Analysis

We received a total of 411 responses to our 100 posted ads. 68 ads received at least one response, and 40 of the 50 pairs of ads received at least one response between them. The responses included 105 offers. 54 of the ads received at least one offer, and 34 of the 50 pairs received at least one offer between them.

All three outcome metrics were higher for the treatment group than the control group. The outcome means for all three metrics are shown in Table 3.

One ad was flagged and removed before it had been up for a full two days, but otherwise the ads were all posted and taken down on schedule in their assigned cities. The ad that was flagged and removed was in the treatment group and received a relatively large number of responses before it was removed. It is reasonable to assume that the more attention an ad receives, the more likely it is to be flagged by Craigslist users, so in this way its removal can be interpreted as a sign of its success. If so, the overall impact of this pattern would be to temper the magnitude of any findings by limiting the response count for some of the more successful ads.

STATISTICAL POWER

We conducted a power analyses of our primary outcome measurement using Stephane Champely's[2] "pwr" package for R. We based our analyses on the paired t-test comparing the response total means in the treatment and control groups, which is equivalent to a regression of the response total on treatment status using pair fixed effects as a covariate.

With 50 pairs we had 80% power to correctly reject the null hypothesis of no effect at 0.05 significance level for a Cohen's d effect size of 0.40. Post hoc, we calculated our observed

[2] Champely, S. "Package 'pwr' - CRAN." 2015. <https://cran.r-project.org/web/packages/pwr/pwr.pdf>

effect size to be 0.15, clearly less than the minimum detectable effect of 0.40. Acknowledging the limited relevance of post hoc power analyses, assuming that the true ATE were reflected by our observations, we would have had only 18% power to correctly reject the null hypothesis. In order to have had 80% power to correctly reject the null hypothesis we would have had to have posted 351 ads.

EXPERIMENTAL RESULTS

Our key estimate is that the treatment increased the number of responses by 0.62 per ad. With a p-value is 0.29 and a 95% confidence interval of [-0.56,1.80], this estimate cannot be taken as statistically significant.

We arrived at the estimate using a number of techniques. A paired t-test and an OLS regression with pair fixed effects as a covariate proved to be equivalent. We also conducted a randomized inference for additional verification, in which we randomized treatment assignment to ad/context subjects, blocking by matched attributes (i.e. ensuring that one member from each pair was assigned to control and the other to treatment). For 10,000 repetitions, the absolute value of the re-randomized estimate had a 0.29 probability of exceeding the observed ATE, confirming the p-value from the t-test and regression.

The response count data is positively skewed with a high proportion of zeros. Figure 1 shows a histogram to the response count. Figure 2 shows a qq-plot of the residuals from the OLS regression, illuminating the potentially misleading character of the linear relationship. Based on these characteristics of the data, and given that the variance (4.33) was similar to the mean (4.11), it was also appropriate to fit the data to a Poisson regression, which yielded an estimate of 0.15 (log response count) and a p-value of 0.13, still not significant.

Both OLS and Poisson regression were run with four covariate sets: 1) without any covariates, 2) with pair fixed effects as described above, 3) with posting city population as an additional covariate, and 4) with fixed effects for city, day and author in addition to population and pair covariates. The results are summarized in Table 4 (OLS) and Table 5 (Poisson). The covariate-free models were not particularly useful other than to give a basis for comparison for the value of the paired design. Population and other fixed effect covariates did not appreciably alter the results.

Finally, we conducted OLS regression tests on the two secondary outcomes, offer count and offer amount, using the same covariate sets. Not surprisingly, given that the number of nonzero observations was less for these outcomes than for the primary outcome, estimates were even farther from being statistically significant. Results are summarized in Table 6 (offer count) and Table 7 (offer amount).

A MACHINE LEARNING APPROACH

As an exercise, we probed the connection between the treatment and outcomes using supervised machine learning. The data were split randomly, with approximately ⅔ put into a training set and ⅓ into a test set. Using the outcomes as features and the treatment status as labels, we attempted to classify the test data with a model trained on the training data.

In order to eliminate between-pairs variation, we considered the difference between a given observed value and its pair mean as the feature value. We tried various combinations of features ranging from single outcome variables to all three outcome variables with city population and city, posting day and author “fixed effects”. We also tried a number of different machine learning algorithms. The best result was 66% accuracy, achieved with all three outcome variables but no other features, using a logistic regression model.

Given that 50% accuracy in predicting treatment status can be achieved with random classification, a result of 66% cannot be considered significant. Machine learning algorithms generally work best with vastly more data. Were we to allocate any significance to the exploration, at least some additional data would be necessary to use as a final test set in order to exclude overfitting. The machine learning code is available for reference in the project repository.

IV. Discussion and Conclusions

As our experiment showed no statistically significant results, our interpretation must by necessity be limited. We are unable to credibly speculate on the drivers of variation in the number of responses in our treatment versus our control groups. It is probable that, were we to rerun the experiment, we would receive substantially different results due to sheer randomness.

We attempted to support our assumptions by randomizing the combinations of cities, posting days and ad authors. That fixed effects for these variables did not appreciably improve our regression models might be taken as an indication that we were successful in doing so, but heterogeneity in these variables weakens any potential conclusions and generalizability. We were somewhat surprised that including population as a covariate did not strengthen the regression models, given that the populations of the cities ranged from around 2 million to around 18 million. Again, successful randomization may have played a part here, but there may be other things at play. For example, supply may be correspondingly greater in larger cities, mitigating additional demand. It would be possible to assess this in future experiments by doing an inventory of comparable products available on Craigslist or in other classifieds. It is also possible that geographical spread has at least as much impact on customer interest as population alone.

To improve the experiment in the future, we would take several deliberate steps. First, we would seek out data on the actual purchasing habits of Craigslist shoppers, perhaps by contacting and securing the approval of Craigslist itself. In our current experiment we used regional populations as a proxy for the size of the Craigslist shopping population, but we were unable to estimate the number of actual shoppers from within that population. We ideally would have a large number of covariates, such as the popularity of certain products on craigslist; the incomes of the average craigslist buyer per city, and so forth. Due to the limited time allowed for the experiment, and to avoid “fishing expeditions” we do not include any other post hoc variables beyond the city population.

Second, we would carefully control the days on which our postings occurred in order to avoid time-dependent effects, such as those from Thanksgiving. Similarly, perhaps our experiment would have been more powerful if we ran the treatment and control ads in the same city across time, rather than at the same time across cities.

Third, we would have increased the number of advertisements we ran per the statistical power calculations discussed above.

Fourth, we would implement a process for assessing the consistency and effectiveness of the treatment across ads and ad authors, for example, by having independent reviewers fill out a questionnaire on the characteristics of the ad copy.

While no significant conclusions can be drawn, the data we gathered weakly point towards an improved customer response to ads with that contain first-person language. A small effect size here can translate into a significant gain, especially for those posting classified ads regularly. An extra response on average for every couple of ads and an additional 5%-10% on offers would be a desirable return on the minimal investment required to adjust ad copy. Further study may therefore be warranted.

References

Jennifer L. Doleac & Luke C.D. Stein, 2013.

"The Visible Hand: Race and Online Market Outcomes," Economic Journal, Royal Economic Society, vol. 123(11), pages F469-F492, November.

Gerber, Alan, and Green, Donald (2012). *Field Experiments, Design, Analysis and Interpretation.* New York: W.W. Norton & Co.

All project data and resources are available in github:

https://github.com/kyleiwaniec/w241Project

Tables

Table 1: Allocation of treatment to city from the random assignment

	City 1	City 2	City 3	City 4	City 5
Population (millions):	2.1	7	3.25	8.4	18.55
Treatment Ads:	10	10	10	10	10
Control Ads:	10	10	10	10	10
Author 1 Ads:	3	2	3	7	5
Author 2 Ads:	5	7	3	3	2
Author 3 Ads:	7	5	2	3	3
Author 4 Ads:	2	3	7	1	7
Author 5 Ads:	3	3	5	6	3

Table 2: Allocation of treatment to posting day from the random assignment

	Day 1	Day 2	Day 3	Day 4
Treatment Ads:	15	11	9	15
Control Ads:	10	14	16	10
Author 1 Ads:	5	8	3	4
Author 2 Ads:	3	4	4	9
Author 3 Ads:	8	6	2	4
Author 4 Ads:	6	4	7	3
Author 5 Ads:	3	3	9	5

Table 3: Summary of Outcomes

	Treatment	Control
Mean Response Count:	4.42 (0.86)	3.8 (0.81)
Mean Offer Count:	1.06 (0.24)	1.04 (0.22)
Mean Mean Offer:	222.25 (0.41)	206.58 (0.33)

Table 4: Total Responses - OLS

	Dependent variable:			
	rtotal			
	(1)	(2)	(3)	(4)
treatment	0.620	0.620	0.620	0.666
	(1.178)	(0.585)	(0.590)	(0.587)
population			-0.028	0.064
			(0.066)	(0.074)
Constant	3.800***	3.690*	3.838*	3.747
	(0.833)	(2.088)	(2.134)	(2.463)
Pair Fixed effects?	No	Yes	Yes	Yes
Day Fixed effects?	No	No	No	Yes
Author Fixed effects?	No	No	No	Yes
Observations	100	100	100	100
R^2	0.003	0.877	0.878	0.903
Adjusted R^2	-0.007	0.752	0.748	0.771
Residual Std. Error	5.890 (df = 98)	2.924 (df = 49)	2.949 (df = 48)	2.806 (df = 42)
F Statistic	0.277 (df = 1; 98)	6.997*** (df = 50; 49)	6.749*** (df = 51; 48)	6.861*** (df = 57; 42)
Note:				**p<0.1;* ***p<0.05;* ***p<0.01

Column 1 is the regression of treatment on total responses. Column 2 includes the ad pair fixed effects. Column 3 includes ad pair fixed effects and population. Column 4 includes all fixed effects: ad pair, population, city, day and author. None of the results are statistically significant. Adding ad pair fixed effects reduces the Standard Errors. But adding other covariates does not add any value.

Table 5: Total Responses - Poisson

	Dependent variable:			
	rtotal			
	(1)	(2)	(3)	(4)
treatment	0.151	0.151	0.138	0.134
	(0.099)	(0.099)	(0.103)	(0.115)
population			-0.007	0.008
			(0.016)	(0.018)
Constant	1.335^{***}	1.308^{***}	1.352^{***}	1.497^{***}
	(0.073)	(0.358)	(0.372)	(0.495)
Pair Fixed effects?	No	Yes	Yes	Yes
Day Fixed effects?	No	No	No	Yes
Author Fixed effects?	No	No	No	Yes
Observations	100	100	100	100
Log Likelihood	-435.057	-152.203	-152.112	-141.164
Akaike Inf. Crit.	874.113	406.406	408.223	398.329
Note:				$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

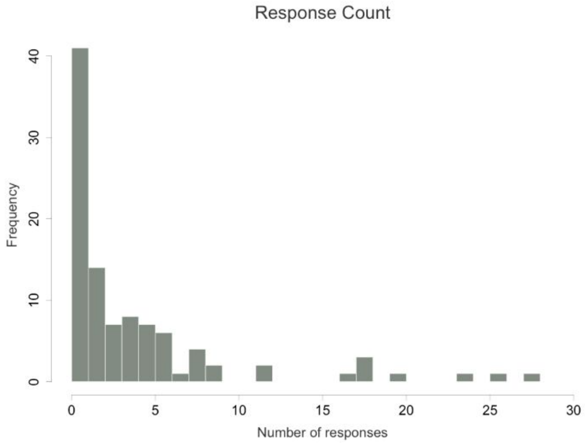
Table 6: Number of Offers - OLS

	Dependent variable:			
	roffer			
	(1)	(2)	(3)	(4)
treatment	0.020	0.020	0.020	0.022
	(0.324)	(0.279)	(0.282)	(0.259)
population			-0.008	0.025
			(0.032)	(0.033)
Constant	1.040***	1.490	1.530	1.819
	(0.229)	(0.997)	(1.020)	(1.087)
Pair Fixed effects?	No	Yes	Yes	Yes
Day Fixed effects?	No	No	No	Yes
Author Fixed effects?	No	No	No	Yes
Observations	100	100	100	100
R^2	0.00004	0.628	0.629	0.749
Adjusted R^2	-0.010	0.249	0.234	0.408
Residual Std. Error	1.619 (df = 98)	1.396 (df = 49)	1.410 (df = 48)	1.239 (df = 42)
F Statistic	0.004 (df = 1; 98)	1.655** (df = 50; 49)	1.593* (df = 51; 48)	2.199*** (df = 57; 42)
Note:				$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 7: Average Offer Amount - OLS

	Dependent variable:			
	avgoffer			
	(1)	(2)	(3)	(4)
treatment	15.67	0.91	4.91	2.12
	(50.83)	(20.16)	(22.06)	(16.35)
population			2.31	3.65
			(4.09)	(4.06)
Constant	206.58***	237.05***	222.90***	292.57**
	(34.27)	(34.92)	(43.99)	(55.13)
Pair Fixed effects?	No	Yes	Yes	Yes
Day Fixed effects?	No	No	No	Yes
Author Fixed effects?	No	No	No	Yes
Observations	44	44	44	44
R^2	0.002	0.98	0.98	1.00
Adjusted R^2	-0.02	0.92	0.91	0.96
Residual Std. Error	167.89 (df = 42)	47.29 (df = 10)	48.98 (df = 9)	32.59 (df = 3)
F Statistic	0.09 (df = 1; 42)	15.78*** (df = 33; 10)	14.28*** (df = 34; 9)	27.85*** (df = 40; 3)
Note:				**p<0.1;* *****p<0.05;*** ***p<0.01

Figures



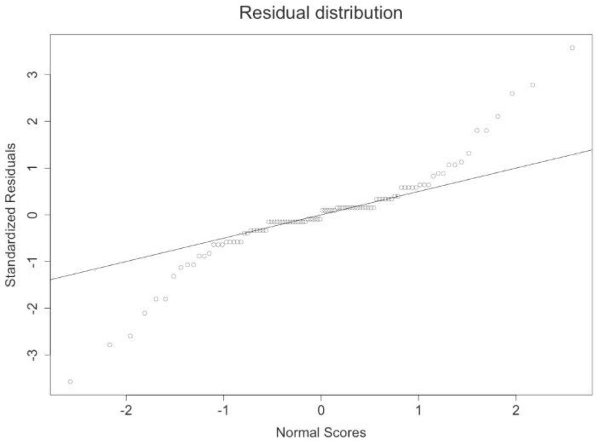


Figure 1. Distribution of responses to ads

Figure 2. qqplot