Teespring, Addepic, and VomitStain

By Kevin Dugan, Jason Sanchez, and Manish Sannat

A randomized experiment evaluating the public appeal of startup companies based on their names yields significantly negative results for especially poor company names. Additionally, outcomes for companies with randomly generated names trended negatively, but in most cases the value was not statistically significant at the .05 level. We draw the conclusion that making the effort to select a quality name for a company is likely time well spent. Quantifying the cost of time spent choosing a name is outside the scope of this experiment; we simply investigate whether the name influences public interest. We also discover that the power of funding support can help a company overcome a poor name.

Introduction

The process of establishing a commercially appealing name for a fresh startup company can be one of many frustrating activities of entrepreneurship. Choosing an inappropriate name, such as *VomitStain,* might invoke negative connotations toward the organization. Whereas a more descriptively appropriate name, such as *Teespring*, may be more intriguing to potential users. Along the same lines, startup names that are neither obtuse nor perfectly relatable to common lexicon, such as *Twitter*, can be just as successful (or not) as organizations with either good or terrible names. Think back to the first time you heard the term *Twitter* - did that make you think of a new way to use social media? Probably not. Regardless, *Twitter* was successful and the associated vernacular (‘tweets,’ ‘followers,’ etc) has become embedded in the everyday lexicon.

We study this particular curiosity in order to identify if an organization’s name plays a key role in the public perception of the company. Regardless of how long the team spends deciding on a name, that time is money. What happens when the team achieves the perfect name after two weeks, only to find out that all the acceptable variations of domain names have been bought up and are not for sale? What if the name has legal roadblocks? Do you start the process all over again, or just hastily create an acceptable alternative and move on? We offer experimental evidence that informs how much influence a startup company's name may, or may not, have on public interest. After all, why spend an inordinate amount of time on choosing a name, if the name is not really what drives success or failure?

A. Layout of Remainder of Document

The remainder of this document is organized as follows. Following this introduction, Section I provides details on the experiment's design. This includes how the experiment was developed and presented to subjects, the hypotheses being examined, outcomes of most interest, organization of factorial design, and an explanation of pilot studies and power calculations. Section II focuses on the results of the experiment. This section details the statistical analysis of the experiment - average treatment effects (ATEs), model summaries, statistically significant results, and analysis of covariate interactions. Also, descriptive assessments of survey respondents will be provided. This paper will conclude with final assessments and lessons learned. An appendix is provided with detailed figures and tables.

I. Experimental Design

The experiment was organized to measure the tradeoff between how much time is invested in coming up with the name of a company and the success of the company. Ideally, we would randomly assign how much time thousands of new startups have to determine their name and measure the corresponding impact on profitability (or some other key metric). Unfortunately, our budget for this project was not big enough for us to run such an ideal experiment.

Instead, we carefully crafted variables we thought would proxy the variables we were interested in. Instead of a profitability metric, we had survey participants read a company description and rate the company on a scale of 0 to 10 based on whether or not they would recommend we invest in the company. The purpose of this was not to see whether an investor would invest in the company; it was to simplify the rating process for the survey respondent.

A. General Structure

We first decided what information to present about a company to each subject. Six actual startup companies were selected by the research team from AngelList. We presumed the *actual* names of the startups were ones the founders of the company put a lot of effort and time into selecting.

In addition to the actual name of the company, two fictitious names were derived for each company, for a total of 18 names. Each name was placed into one of three categories: actual, nonsense, and terrible. Table 1 shows a complete listing of company names used in the experiment.

Nonsense names were generated from a nonsense word generator, where each team member generated a list of 50 nonsense words for another team member (see appendix for an example of this). Each of us picked a few nonsense words from these lists to be alternative names for the startups. Nonsense names served as names for which very little (if any) time at all was spent deciding the name.

Terrible names were created by each team member with the aim of making the name as unappealing as possible without being obscene. The objective was to set a lower bound where we could detect a definitive negative response.

Table 1. Startup Company Names by Category

3 Categories 6 Companies 18 Total Names	**Company**					
	1	**2**	**3**	**4**	**5**	**6**
Actual	Appboy	Teespring	Lystable	LendUp	BitGym	Homey
Nonsense	Serefly	Addepic	Sackle	Grativie	Hacklow	Bandity
Terrible	CatRash	VomitStain	Tablebum	LoanShark	TortureTour	DoodyCalls

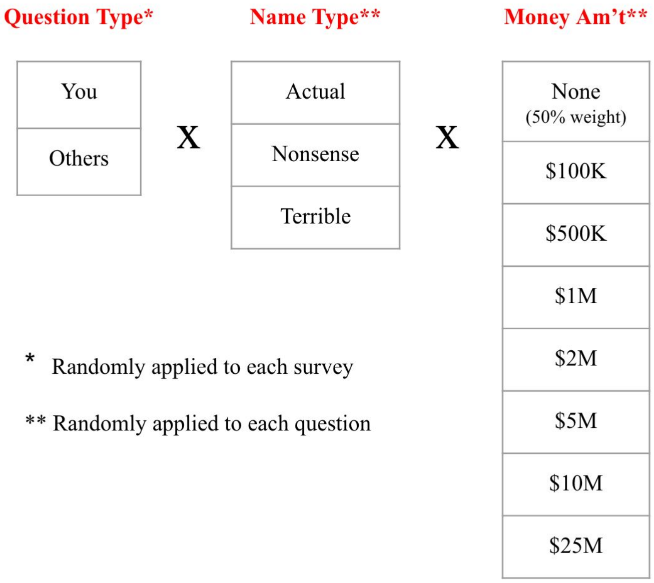
For each company, regardless of the name shown, the company's actual description and logo were presented for each question. There were two additional random pieces of information provided to respondents. First was an amount of seed money already invested in the company, and second was a randomization of phrasing for the question posed to the participant.

We chose eight different amounts of money to present as seed money. The seed money would be 'None' for 50% of the questions, and the remaining seven amounts were randomly assigned with equal weighting. The other covariate involved varying the phrasing of the question posed to the subject. Either the subject was asked to what degree *they* would recommend we invest, or to what degree they thought *others* would recommend we invest. The purpose of this question was to test whether or not perceived individual biases impacted the ratings given to companies.

The survey was carefully crafted to capture the explanatory variables and responses in a fair manner. The core of the survey consisted of rating each of the six companies. The order of the companies was shuffled for each participant. Additionally, each survey was randomly placed into one of two categories based on whether the participant was to rate the companies from his or her own perspective or was to rate companies based on how the average person would be expected to rate them. For each question, one of the three name types was randomly presented and a randomized amount of seed money (from eight possibilities) was selected. This resulted in a 2 x 3 x 8 factorial design for each question. Each survey question would either be in the control group or treatment group, depending on the randomized names. A question was considered as being in the control group if the actual name of the company was used and in the treatment group otherwise. See Figure 1 for a graphical representation. The survey design was complex enough that the team had to build a custom service using AWS Lambda to serve our specific randomization scheme to Qualtrics.

Following the six core questions, demographic data was collected about each subject. Subjects were provided the options to choose from nine age ranges, three levels of education, and five categories asking the subject to self-rate their perception of their own business acumen. Subjects were also asked to venture a guess as to what we would do with their answers in order to help the team understand how the survey was being perceived. Please see the appendix for a copy of the survey.

Figure 1. 2 x 3 x 8 Factorial Design



* Randomly applied to each survey

** Randomly applied to each question

B. Hypotheses and Power Calculations

The hypothesis being evaluated is that the name of a company will not matter - that is, there will be no difference in the average rating between actual and nonsense names or between actual and terrible names. More specifically:

$$H_{0(1)}: \textit{average actual name ratings} = \textit{average nonsense name ratings}$$

$$H_{a(1)}: \textit{average actual name ratings} \neq \textit{average nonsense name ratings}$$

$$H_{0(2)}: \textit{average actual name ratings} = \textit{average terrible name ratings}$$

$$H_{a(2)}: \textit{average actual name ratings} \neq \textit{average terrible name ratings}$$

The great name, in this case, was considered to be the actual name of the company. By using actual startup company names in the experiment, and then using fictitious names as the intervention, we strive to observe how changes in the names reflect in public interest levels. The covariates of seed money and question phrasing were treated as interactions with the name category. To measure the causal effect of the name on the public's perception of the company, we grouped the ratings according to the *category* of the name (actual, nonsense, terrible).

With respect to sample size and power calculation, the objective was to be able to detect an average treatment effect of $\pm .5$ on the rating scale, with 95% confidence level 80% of the time. The power calculations established that we would require a minimum of 130 valid surveys or 780 rating data points. The pooled standard deviation was calculated using the variance of the ratings across funding levels from the pilot study data for the control and treatment groups. Sample sizes were calculated using the standard benchmark of 80% for statistical power. We averaged the power study results of 1,000 simulations of the experiment, sampling with replacement from the pilot study data. Once we achieved an average statistical power of 80%, we used that sample size as the baseline minimum.

Given that our pilot study was significantly less than 1,000 ratings, and that we conducted sampling with replacement, the simulations had multiple records of the same individual survey.

C. Participant Selection

As mentioned earlier, Mechanical Turk was used to recruit participants for the experiment. While a convenient way to reach willing participants, Mechanical Turk can lead to violations of the exclusion restriction. We ran multiple pilot studies using Mechanical Turk, and on each run, we increased the risk that a subject would participate in the survey more than once. Based the experiment's design, we ran a greater risk that a participant could have experienced both control and treatment. If a subject took the survey twice, they could have experienced a control name and a treatment name for the same company(s) in their two separate attempts. Post-survey, we did pose a question to the subject that asked whether they had taken the survey before (with no penalty for payment if they answered 'yes'), and then excluded those positive responses from our final dataset.

The assumption of non-interference is also an issue when using Mechanical Turk. The assumption here is that one subject taking the survey does not impact another subject. The Turkopticon community provides participants a way to convey information about a particular task. There could have been an instance where information about the task that could affect others' responses would have been posted. For example, if we received negative feedback from the Turker community, that could have affected the quality of our participant responses. At this point, we are unaware of any negative postings, and had received at least two instances of positive postings.

There were several survey designs we tested before settling on our final design. For example, we attempted to frame the survey in such a way so we could use a difference-in-differences model. We started by asking the participants to rate the six startups. Then we asked them several demographics questions. Then we asked them to rate the same six startups, but we shuffled the order and would change the names randomly as well. Our hope was that the participants would forget enough about the first set of startups so that they would rate the second set as if they had not seen the first set. If this happened, we would be able to build a difference-in-differences model which would dramatically reduce the variance that comes from differences in individual participants. Unfortunately, many participants caught onto this ruse.

A primary concern of ours was data quality. Mechanical Turk is notorious for poor survey answers due to people attempting to complete the surveys as fast as possible or for being uneducated or unfamiliar with US cultural views. To combat fast survey design we included a final question at the end that asked the participant to select from a list of 18 possible startup names the 6 startups referenced in the survey. They had to get at least 3 out of 6 choices correct for us to use their ratings in our analysis. Additionally, the fastest and slowest 5% of respondents were eliminated to eliminate people who rapidly clicked through the survey or who took an unreasonably long time completing the survey.

To help ensure we collected responses that were from qualified applicants, we asked three prescreen questions:

1. If the day before yesterday was Monday, what is today?
2. How many feet are in a yard?
3. Which of the following is least likely to be the next president of the United States? {Clinton, Trump, Sanders, Obama}

The first question tested English comprehension skills, the second question tested a cultural norm, and the third question tested current US knowledge. A surprisingly large percentage (20%) of respondents failed this prescreen even though we required participants to be from the US.

II. Experiment Results

We wanted to only use the surveys that met the prerequisite level of completion for evaluating the experiment's results. Of the 560 surveys taken, we removed the following records:

1. Failing qualifier question responses.
2. Rating values were missing
3. 5% slowest and 5% fastest respondents
4. Respondents that could not recall at least 50% of the startup names presented to them.
5. Participant who had taken the survey more than once.

In all, we had 286 valid surveys, which translated to 1,716 rating data points. Of the 1,716 ratings, 550 (32%) were for actual names, 565 (33%) ratings were for nonsense / random names, and 601 (35%) ratings were for terrible names. The distribution of names was relatively even. Power calculations estimated needing 130 valid surveys and we were able to receive more than twice that amount.

Keep in mind, there are six ratings provided per individual. The assumption we make is that each startup is rated independently of the other startups (non-interference between questions). That is, a person's rating of one startup company will not influence their rating on a completely different startup. This assumption was held for the power calculations as well.

Our covariate measurements, particularly the estimates for actual names of the startups, provided insight as to how the respondents generally perceived the company. The respondents did not know about the concept of the name categories we used for our analysis. Therefore, a statistically significant effect on an actual startup's name tells us that particular startup was perceived strongly in a particular direction.

A. Average Treatment Effects for Outcomes of Interest

We estimate two average treatment effects (ATE) in this experiment - given there are two levels of treatment in the design. We are interested in the causal effects of having a *nonsense* name and also a *terrible* name.

Let $Y_i[0]$ represent the ratings outcomes for *actual* names.

Let $Y_i[1]$ represent the ratings for *nonsense* names.

Let $Y_i[2]$ represent the ratings for *terrible* names.

The equations for deriving the ATE's for each outcome of interest are as follows:

(1) $$ATE_1 \equiv \frac{1}{N}\sum_{i=1}^{N}(Y_i[1] - Y_i[0]) .$$

(2) $$ATE_2 \equiv \frac{1}{N}\sum_{i=1}^{N}(Y_i[2] - Y_i[0]) .$$

The average rating for the *actual* names is 4.88, *nonsense* names is 4.61, and 4.36 for *terrible* names. Therefore, $ATE_1 = 4.61 - 4.88 = -0.26$ and $ATE_2 = 4.36 - 4.88 = -0.52$. We can confirm our ATE's, and obtain significance levels easily using a linear model for the two treatments. Results of the linear model listed in Table 2 below.

(3) $$y_i = \beta_0 + \beta_{1i}nonsense + \beta_{2i}terrible + \varepsilon_i .$$

Naive Model.

Table 2. Average Treatment Effects for Startup Name Categories

	Estimate	**Std Error**
Actual Name (Intercept)	4.878 ***	0.114
Nonsense Name	-0.264 .	0.160
Terrible Name	-0.520 ***	0.158

Significance Codes: '***' .001 '.' 0.1

From the basic model, we observe the nonsense name is not statistically significant at the .05 level. Therefore, we fail to reject the null hypothesis. Being statistically significant at the .1 level indicates directionally how nonsense names are perceived. The terrible name estimate is statistically significant at the .05 level. We reject the null hypothesis and conclude that average ratings for actual names are consistently .5 higher than the average terrible name ratings. As expected, terrible names do negatively impact ratings of a startup company. These preliminary results confirm that we were able to devise bad enough names to invoke a negative response.

One of the hidden agendas of this study was to test something that a certain team member strongly believed to be true (cough, cough, Jason). Namely, spending time coming up with a "perfect" startup name does not materially help a company over just quickly picking a reasonable name. We were not content with just proclaiming that there is no statistically significant difference between the nonsense name and the actual name, so we decided to go a few levels deeper and add covariates.

B. Covariate Analysis

Several covariates were included in the experiment to potentially account for impacts of factors outside of the name categories. This, in turn, improves the precision of our estimated ATEs. We will address the covariates in groups and discuss results of models that build upon one another by the inclusion of additional covariates for each model. We will discuss only some of the findings of the models specifically in this section. A full schedule of outcomes is located in Table 3 in the appendix.

Model 1

The first set of covariates up for discussion are the individual names of the startups we selected. Remember, the outcome of interest were the ATEs of the *categories* of the names, not the names themselves. It is possible that some startups may be perceived either positively or negatively naturally based on the company's business model or other factors, regardless of the name. The results of a linear model that used indicator variables for the actual startups revealed that three of the companies had statistically significant estimates.

$$y_i = \beta_0 + \beta_{1i} nonsense + \beta_{2i} terrible + \beta_{3i} startup1 + \beta_{4i} startup2 + ... + \varepsilon_i \tag{4}$$

Model 1. Model with Startup Name Covariates

The statistically significant estimate for the startup named *Homey* was almost a full point below the average rating across all startups (-0.983 (.219)). This company was not well-received by our participants. Looking at the average ratings for 'Homey' by name category, we see that despite the name, the ratings are generally the same, low average rating. In fact, the average rating for the terrible name is 0.03 **higher** than the actual name. Was the terrible name not terrible enough? Was it the fact that our participant audience was a crowd that enjoys puns (*DoodyCalls*)? With consistently low scores across the board, *Homey* ratings give us the assurance that each name category received the same general effect from that startup.

As a thought exercise, why would *Homey* be rated so much lower by our respondents? Consider the demographics of our audience - a generally younger crowd either in college or perhaps having recently graduated college. The business model behind *Homey* is that it is a mobile app that lets members of families take pictures of a pile of dirty laundry and assign it to another family member as a chore. With a generally younger demographic, our respondents may not be the target audience for that startup. Therefore, the company may seem unappealing on that basis alone.

Table 3. Average Rating by Startup

	Average Rating of Startups					
	1	**2**	**3**	**4**	**5**	**6**
Actual	Appboy 4.94	Teespring 5.26	Lystable 5.32	LendUp 5.05	BitGym 5.06	Homey 3.66
Nonsense	Serefly 4.2	Addepic 4.72	Sackle 5.35	Grativie 5.4	Hacklow 4.32	Bandity 3.62
Terrible	CatRash 4.21	VomitStain 3.32	Tablebum 5.32	LoanShark 5.01	TortureTour 4.55	DoodyCalls 3.69

Consider two more statistically significant startups that were observed, but in a positive direction. *Lystable* (0.686 (0.219)) and *Lendup* (0.515 (.219)) startups were generally well-received by our respondents. Regardless of the name shown for these two startups, they both had a statistically significant average rating that was greater than the baseline average rating. With *Lystable*, there is essentially no difference between the ratings of the name categories. The *Lendup* nonsense name actually fared better than the actual name, however the average ratings were consistent across all name categories. Like *Homey,* this indicates that effects on each name categories were even.

This is an important point because we have three startups where the name does not appear to matter at all. Then we have three startups where the actual names do have the highest ratings. We presume the consistent average ratings for *Homey, Lystable,* and *Lendup* are a result of a business model that would commonly be expected to do well or poorly. With the remaining three, the ratings degrade steadily by name category, which reinforces the estimates calculated for name category in the models.

With respect to our name categories, we can observe how our estimates have changed ever so slightly in this model. The average name rating increased to 4.910 (0.181), a +0.03 effect. The estimates for the nonsense names changed to -0.276 (0.157) from -0.264, with no change in statistical significance. The estimate for the terrible name changed to -0.517 (0.155) from -0.520, and retained the same level of statistical significance as in the original model.

Model 2

We turn our attention to another group of covariates - the demographic data collected from the subjects and the individual of focus for the question, *you* vs. *others.* Since we provided an age range to the participants, in order to scale it for measurement, we took the mean of each age range. Along with age, we considered education level and subjects' self-ratings of their business knowledge.

We observe one covariate of significance - those instances where the subject stated to know 'a lot' about business. The estimate is 0.724 (0.270) and is significant at the .001 level. This indicates that when a person claims to know 'a lot' about business, there are generally increased ratings. The people that identify in this business knowledge category tend to give higher ratings than those that identify in a different business knowledge category. The mean rating for this group is 5.12, a .24 average increase from the original estimate.

We observe our baseline estimate remains relatively the same, 4.83 (0.742) with the standard error increasing by almost 0.6. The nonsense name remains statistically insignificant at the 0.05 level with an increased estimate of -0.290 (0.157). The terrible name estimate increases to -0.533 (0.155) with no change to statistical significance.

$$(5) \qquad Model(1) + \beta_{7i}age + \beta_{8i}focus + \beta_{9i}education.level + \beta_{10i}business.knowledge + \varepsilon_i$$

Model 2. Model with Covariates, No Funding

Model 3

Recall that we included randomized funding information as a part of each company's description. We included those into our running model as an interaction term paired with both the nonsense and terrible name categories. The funding levels were rescaled by taking the natural log of the funding value, and also the funding value in US dollars. The model showed statistical significance for the natural log of the funding value. The estimate is .079 (.0179) and this is an intriguing addition to the model. By taking the natural logarithm of funding levels, we see an estimate that can be interpreted as a percent increase. Multiplying that percent increase by the funding amount provides the change in rating along our scale. What this seems to indicate is that even if the startup is named *VomitStain*, if it had $25M in funding already, the average rating would increase by 1.27 points. Factor in the negative impact of the terrible name and there is still an overall positive effect.

Table 4. Effect of Funding on Ratings

Funding Level	Average Change in Rating
$0	0
$100,000	.912
$500,000	1.04
$1,000,000	1.09
$2,000,000	1.14
$5,000,000	1.21
$10,000,000	1.25
$25,000,000	1.27

The inclusion of the funding caused our outcomes of interest to increase. The estimate for the nonsense name estimate moved to -0.339 (.211) but is not statistically significant. The estimate for the terrible names also increased to -0.630 (.210) and retained its statistical significance.

$$Model(2)\ +\ \beta_{11i}nonsense * fund\ + \beta_{12i}terrible * fund\ + \beta_{13i}fund + \varepsilon_i \tag{6}$$

Model 3. Full Model with Covariates

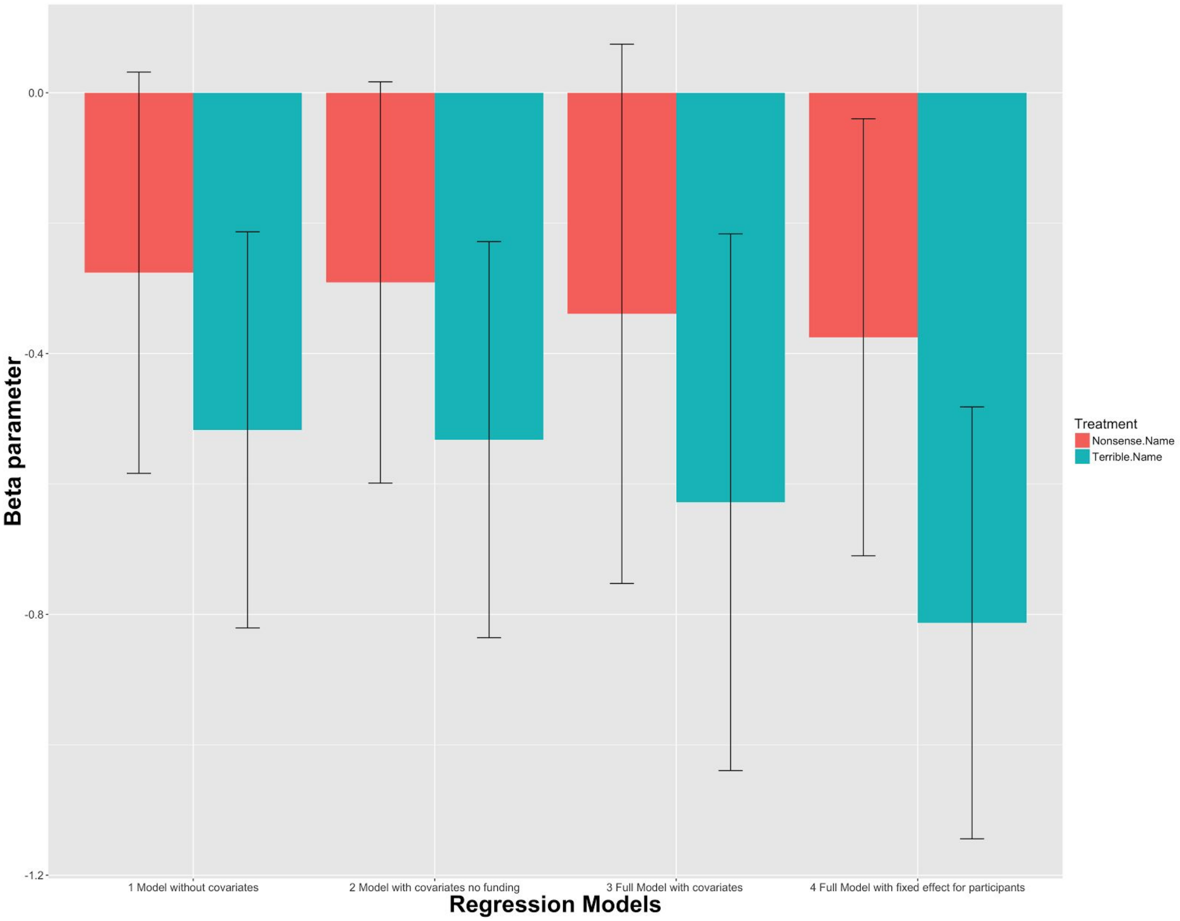
Model 4

As a final model, we included the fixed effects for participants in the form of the unique random codes assigned at the successful completion of the survey. The estimates for the nonsense names increased to -0.404 (0.216), with a p-value of 0.062. While not significant at the 0.05 level, it does indicate strong trending to a reduction in rating. The terrible names estimate grew to -0.744 (.0216), with no change to statistical significance. As we had wanted, we saw that the terrible name was consistently receiving poorer ratings. Figure 2 displays the change in the beta parameters for the nonsense name and terrible name over each model. As the covariates are added, we detect continually falling estimates for nonsense and terrible name categories.

$$Model(3)\ +\ \beta_{14i} random\ code + \varepsilon_i \tag{7}$$

Model 4. Full Model with Fixed Effects for Participants

Figure 2. Beta Parameters by Regression Model



III. Conclusions and Summary

The name of a startup company matters to the public. We have shown that a very poorly named company will receive poor ratings of public interest. Nonsense / random names, while not as statistically significant, still trend strongly toward receiving poor ratings. However, if a very poorly named company, *VomitStain*, happened to have $25M in the bank at the time of the survey, they could overcome the name and receive higher ratings. This experiment cannot answer how *VomitStain* managed to acquire such funding, but we could estimate that *VomitStain* would likely receive higher than average ratings if the money existed.

We observed that poor business models (*Homey*) tend to rate low regardless of their name. Also, good business models (*Lystable, Lendup*) are consistently rated higher regardless of their name. In other startups, as the name degrades, so does public interest. This experiment has shown that at least some amount of time should be dedicated to selecting an appropriate name for a startup company - especially one with a weak business model. The importance of getting funding is a strong indicator that influences public interest, which perhaps could be an item of interest for further experiments.

Key learnings from each team member

Jason

Studies that may seem simple still have a lot of room for unanticipated work. For example, the survey tool did not have the capability to do exactly what we needed. Instead of changing the scope of the project, we built additional capabilities into the tool to give us the randomization scheme we required. Also, it took four different trial tests to get a design that was not flawed. Even then, the data was incredibly messy and required us to throw away almost half of the dataset to get valid results. We also ran into a bug in the Jupyter notebook that gave erroneous results for regressions with higher order polynomial terms and found that the bug did not exist in RStudio.

Experiments actually work. I know this might sound strange, but I am continually shocked that these studies actually work. The participants didn’t know the purpose of the study. We know because we asked them to guess what the purpose of the study was. We listed information on six startups and changed the name of each startup. Participants just looked at the description of each startup and rated them on whatever they felt like. It is understandable how names such as VomitStain might impact the rating and how certain startup ideas were rated higher than others. What is surprising is that there is a detectable (and relatively large) difference between the actual startup name and the nonsense names. For example, Appboy vs. Serefly, Teespring vs. Addepic, and BitGym vs. Hacklow all seem like reasonable alternatives, but all had major differences in rating that were discoverable with DOE methods.

Kevin

Complexity can grow quickly. To explain this experiment to someone, which I’ve done a couple of times, it does not sound very complicated. It usually goes something like, “So we’re showing six companies, randomizing the names, and asking people to rate their interest. And we have really terrible names like 'CatRash' and 'VomitStain,' random words, and the actual names. So we're just trying to figure out if the name makes a difference to the public." But this experiment turned out to be sophisticated - at least for my experience level. I'd say that as each day passed, I felt like I understood a new nuance of what we were trying to accomplish. Incorporating funding level covariates and interpreting how they influence a respondent's rating was one example of where it felt like, 'oh, that's pretty cool,' that I may not have realized on my own.

It takes a team. This may be accurate (and obvious) for many things, but in terms of running a complex experiment, a quality team is invaluable. The benefit of having multiple perspectives proved to be a real asset for me. I gained a great deal of very valuable information from teammates with more experience in areas where I was lacking. With respect to the experiment, we each had perceptions about the design and execution that enabled us to reach an optimal solution given the circumstances. Having additional sets of eyes on an experiment trying to prove causation really boosted the quality of our project.

Manish

Promise small, deliver big. Learned from Jason that it's important to keep the scope of school project limited but deliver it like any professional project. Underneath simplicity of the project objectives, lies sophisticated and well thought experiment design and implementation, backed by careful data analysis and validation.

Knowing how to program is important, but what's more important in the project is **to learn applying fundamentals of field experiments**. Coming from programming background, it took a great effort for me and with the help from my teammates, I was able to see how concepts of field experiments actually work and how to interpret regression results. I am personally thankful to Jason for spending long hours patiently explaining nitty gritty of field experiments and for pushing me to treat the school project like my job depends on it.

IV. Appendix

Table 5. Regression Models

	Initial Model	Model 1	Model 2	Model 3	Model 4
Actual Name (Intercept)	4.88 *** (0.114)	4.910 *** (0.181)	4.830 *** (0.742)	4.490 *** (0.723)	-1.770 *** (8.520)
Nonsense Name	-0.264 . (0.160)	-0.276 * (0.157)	-0.290 * (0.157)	-0.339 (0.211)	-0.404 * (0.216)
Terrible Name	-0.520 (0.158)	-0.517 *** (0.155)	-0.533 *** (0.155)	-0.630 *** (0.210)	-0.744 *** (0.216)
Appboy		-0.200 (0.219)	-0.200 (0.219)	-0.130 (0.211)	-0.122 (0.197)
Teespring		-0.196 (0.219)	-0.196 (0.219)	-0.081 (0.211)	-0.069 (0.198)
Lystable		0.686 ** (0.219)	0.686 *** (0.219)	0.783 *** (0.212)	0.795 *** (0.198)
Lendup		0.515 ** (0.219)	0.515 ** (0.219)	0.495 *** (0.211)	0.490 ** (0.197)
Homey		-0.983 *** (0.219)	-0.983 *** (0.219)	-0.894 *** (0.211)	-0.883 *** (0.197)
Focus			0.042 (0.128)	0.035 (0.123)	1.580 (2.120)
Age			0.015 (0.035)	0.005 (0.034)	0.427 (0.508)
I(Age2)			-0.0003 (0.0004)	-0.0001 (0.0004)	-0.004 (0.005)
Graduated College			-0.214 (0.266)	-0.333 (0.257)	-2.230 (3.580)
Some College			-0.238 (0.272)	-0.437 * (0.263)	-2.870 (2.890)
Business			0.122	0.101	-6.030

Knowledge ('A great deal')			(0.327)	(0.316)	(8.490)
Business Knowledge ('A little')			0.024 (0.174)	0.065 (0.169)	-3.420 (2.990)
Business Knowledge ('A lot')			0.724 *** (0.270)	0.682 *** (0.261)	-2.250 (3.640)
Business Knowledge ('A moderate am't)			0.230 (0.192)	0.279 (0.186)	-2.650 (2.080)
Funding in Dollars				-3.05e-09 (1.97e-08)	0.000 (0.000)
Funding (Natural Logarithm)				0.079 *** (0.018)	0.086 *** (0.018)
Nonsense Name * Funding Dollars				3.86e-08 (2.82e-08)	0.000 (0.000)
Terrible Name * Funding Dollars				6.39e-08	0.000 (0.000)
Nonsense Name * ln(Funding)				-0.003 (0.025)	0.007 (0.026)
Terrible Name * ln(Funding)				-0.010 (0.025)	-0.014 (0.025)
Random Codes (Fixed Effects)					286 estimates

Significance Codes: * $p < 0.1$; ** $p < 0.05$, *** $p < 0.01$

Figure 3. Plot of Average Ratings by Category and Startup Name

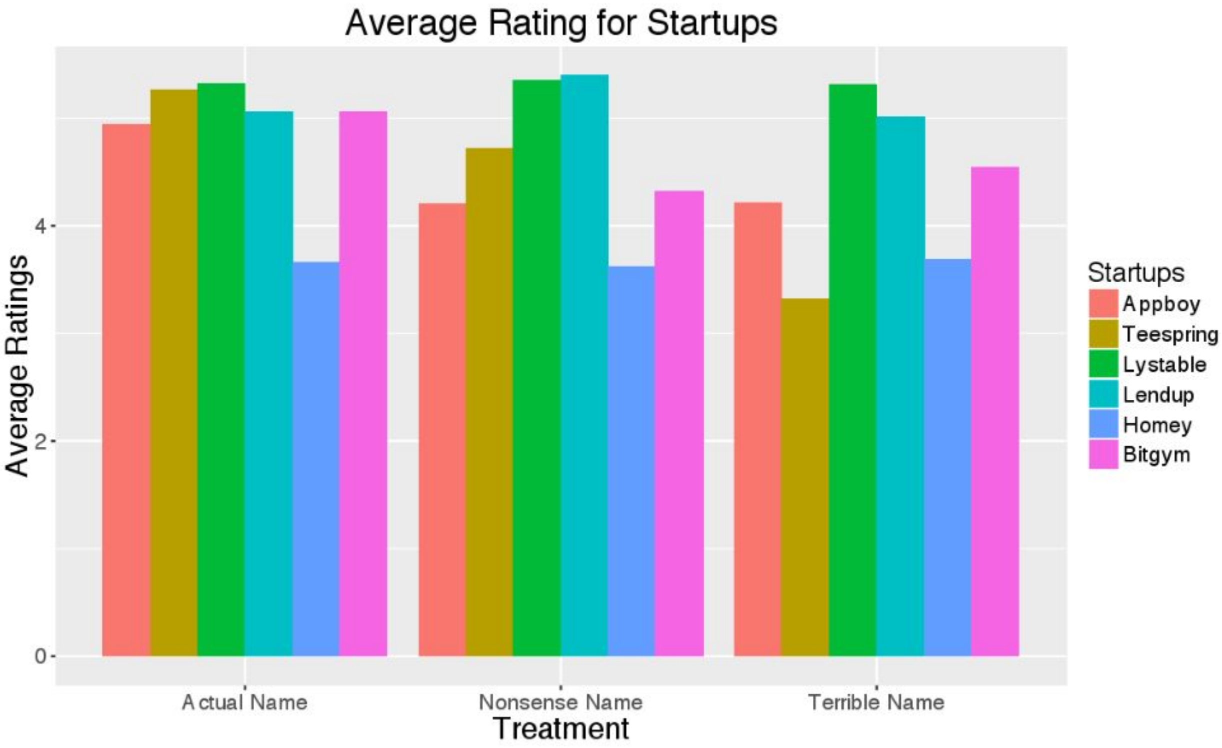


Figure 4. Distributions of Participant Demographic Data

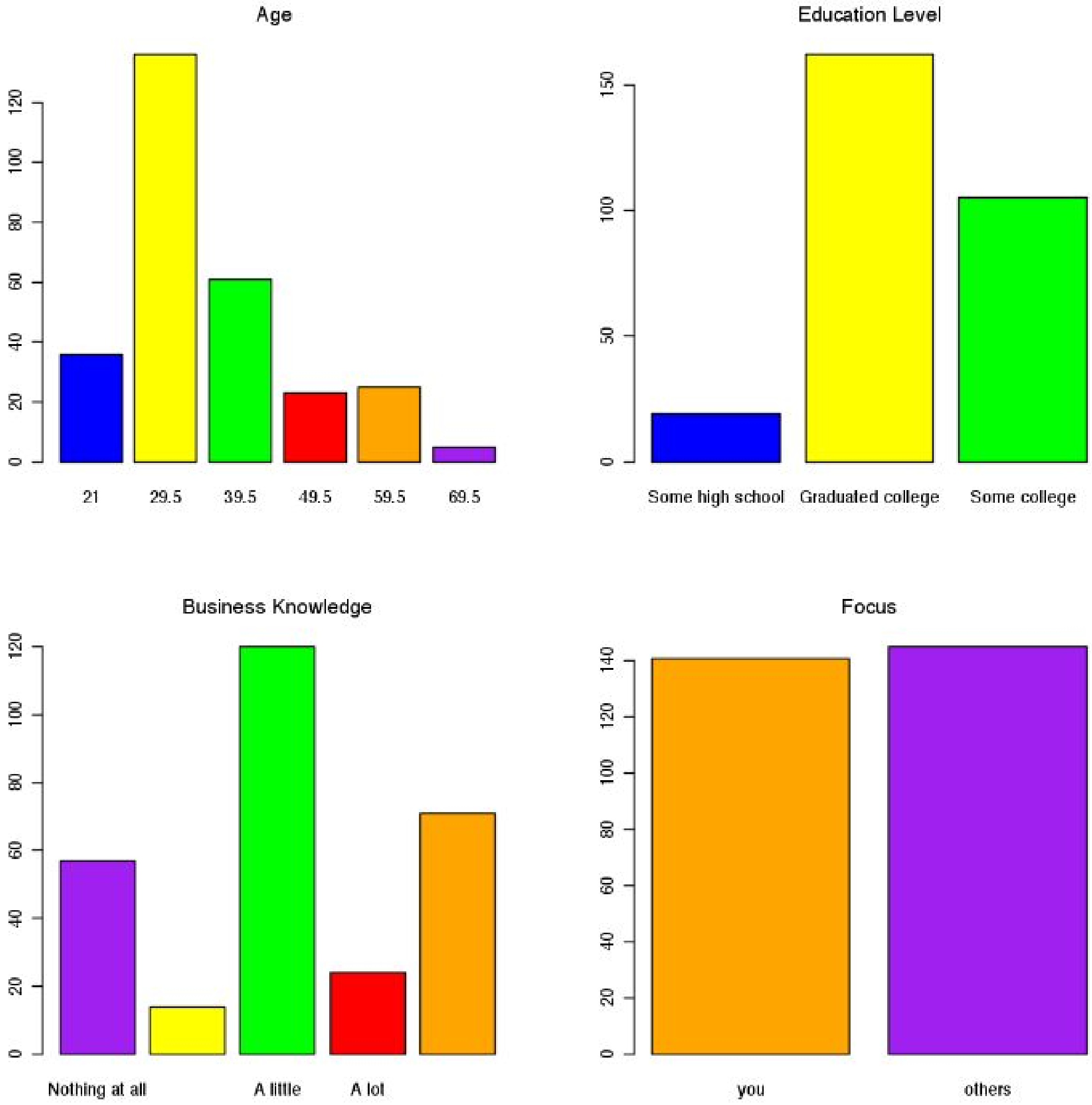
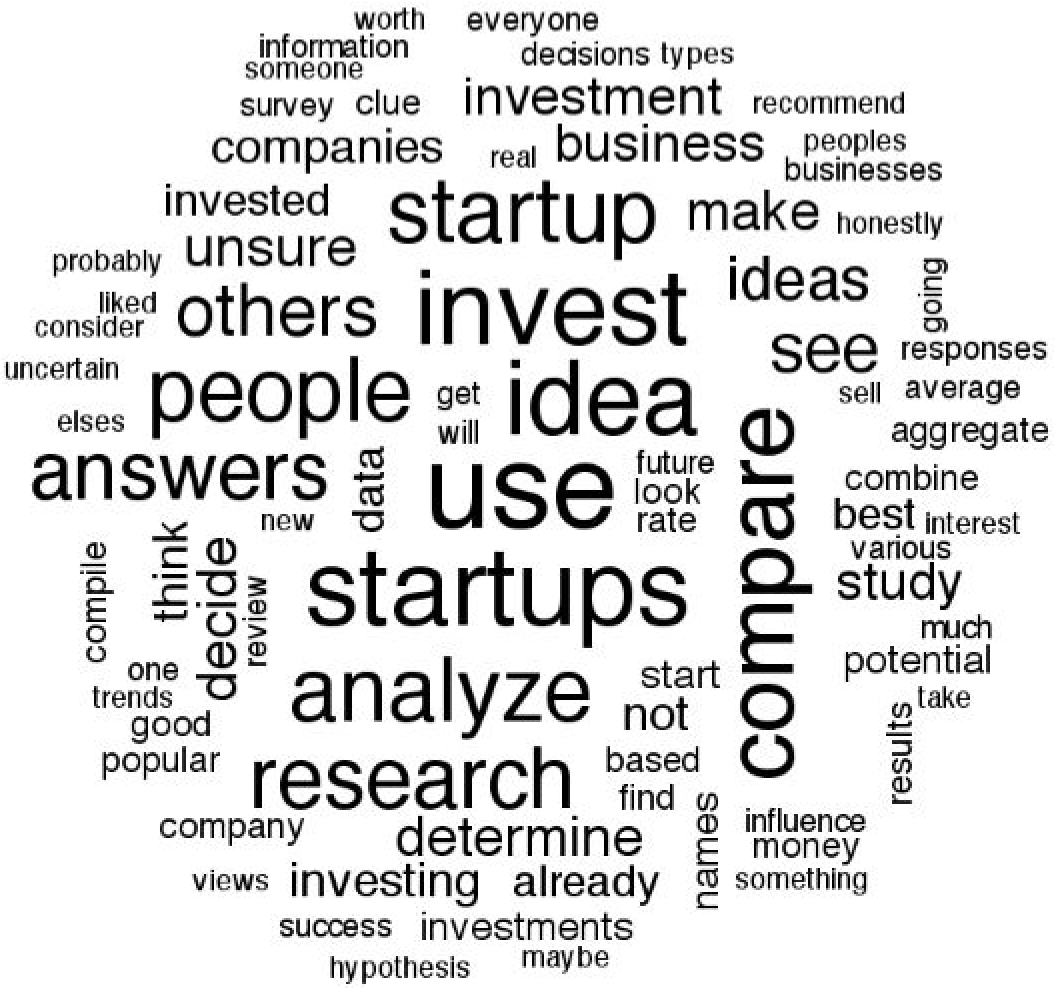


FIGURE 5. Word Cloud of Participant Guesses for Survey Purpose



Survey

W241 - Startup name tests

Q1 The first three questions are required to be answered correctly to continue the survey and to be compensated for your work.

Q2 If the day before yesterday was Monday, what is today?

- Monday (1)
- Tuesday (2)
- Wednesday (3)
- Thursday (4)
- Friday (5)
- Saturday (6)
- Sunday (7)

Q3 How many feet are in a yard?

- 1 (1)
- 2 (2)
- 3 (3)
- 5 (4)
- 10 (5)

Q4 Which of the following is least likely to be the next president of the United States?

- Trump (1)
- Obama (2)
- Clinton (3)
- Sanders (4)

Q9 ${e://Field/wording_description}

Q10 - Appboy ${e://Field/appboy_name}

${e://Field/appboy_name} empowers mobile marketers to build better relationships with their customers. Our robust audience segmentation tool allows companies to use the data from user profiles we generate to create and automate highly personalized marketing campaigns. Thousands of global marketers use us to power over a billion user profiles worldwide.

Amount of funding others have invested in ${e://Field/appboy_name}: ${e://Field/appboy_fund}

Would ${e://Field/wording_you_vs_others} recommend we invest in ${e://Field/appboy_name}?

- 0 (Not at all) (1)
- 1 (2)
- 2 (3)
- 3 (4)
- 4 (5)
- 5 (Maybe) (6)
- 6 (7)
- 7 (8)
- 8 (9)
- 9 (10)
- 10 (Absolutely) (11)

Q11 - Teespring ${e://Field/teespring_name}

${e://Field/teespring_name} allows you to crowdfund custom apparel with no upfront costs, no risks, and no hassle. Design the perfect tee, choose a goal (tipping point) and set a sale price to launch a campaign. Once you reach your goal we handle the printing & shipping and you get a check for the profit.

Amount of money others have invested in ${e://Field/teespring_name}: ${e://Field/teespring_fund}

Would ${e://Field/wording_you_vs_others} recommend we invest in ${e://Field/teespring_name}?

- 0 (Not at all) (1)
- 1 (2)
- 2 (3)
- 3 (4)
- 4 (5)
- 5 (Maybe) (6)
- 6 (7)
- 7 (8)
- 8 (9)
- 9 (10)
- 10 (Absolutely) (11)

Q12 - Lystable ${e://Field/lystable_name}

${e://Field/lystable_name}'s mission is to enable the shift to flexible work, by empowering companies and individuals to work together in new ways. In 2015, we launched a tool that lets companies easily onboard, manage, talk with, and pay their external workers.

Amount of money others have invested in ${e://Field/lystable_name}: ${e://Field/lystable_fund}

Would ${e://Field/wording_you_vs_others} recommend we invest in ${e://Field/lystable_name}?

- 0 (Not at all) (1)

- 1 (2)
- 2 (3)
- 3 (4)
- 4 (5)
- 5 (Maybe) (6)
- 6 (7)
- 7 (8)
- 8 (9)
- 9 (10)
- 10 (Absolutely) (11)

Q13 - LendUp ${e://Field/lendup_name}

${e://Field/lendup_name}'s first product is a socially responsible alternative to payday loans called The Ladder. The Ladder changes the dynamics of the small dollar loan: rather than being a dangerous first step into a cycle of debt, it becomes an opportunity to learn good financial behavior and to build credit through education, gamification, and a transparent fee structure.

Amount of money others have invested in ${e://Field/lendup_name}: ${e://Field/lendup_fund}

Would ${e://Field/wording_you_vs_others} recommend we invest in ${e://Field/lendup_name}?

- 0 (Not at all) (1)
- 1 (2)
- 2 (3)
- 3 (4)
- 4 (5)
- 5 (Maybe) (6)
- 6 (7)
- 7 (8)
- 8 (9)
- 9 (10)
- 10 (Absolutely) (11)

Q14 - Homey
${e://Field/homey_name}

${e://Field/homey_name} is for families with children who are dissatisfied with using sticky notes on their fridge for assigning chores. Our product is a mobile app for convenient management of household chores. You organize tasks by taking pictures of the mess and assigning the work to individual family members.

Amount of money others have invested in ${e://Field/homey_name}: ${e://Field/homey_fund}

Would ${e://Field/wording_you_vs_others} recommend we invest in ${e://Field/homey_name}?

- 0 (Not at all) (1)
- 1 (2)
- 2 (3)
- 3 (4)
- 4 (5)
- 5 (Maybe) (6)
- 6 (7)
- 7 (8)
- 8 (9)
- 9 (10)
- 10 (Absolutely) (11)

Q15 - BitGym ${e://Field/bitgym_name}

${e://Field/bitgym_name} helps you take your cardio workout around the world with a growing library of HD tours which progress at the speed of your exercise. There is no extra hardware to buy or configure because our product uses the front facing camera on smart devices to detect user motions on all types of exercise machines.

Amount of money others have invested in ${e://Field/bitgym_name}: ${e://Field/bitgym_fund}

Would ${e://Field/wording_you_vs_others} recommend we invest in ${e://Field/bitgym_name}?

- 0 (Not at all) (1)
- 1 (2)
- 2 (3)
- 3 (4)
- 4 (5)
- 5 (Maybe) (6)
- 6 (7)
- 7 (8)
- 8 (9)
- 9 (10)
- 10 (Absolutely) (11)

Q6 How old are you?

- Under 18 (1)
- 18 - 24 (2)
- 25 - 34 (3)
- 35 - 44 (4)

- 45 - 54 (5)
- 55 - 64 (6)
- 65 - 74 (7)
- 75 - 84 (8)
- 85 or older (9)

Q7 Education level

- Some high school (1)
- Some college (2)
- Graduated college (3)

Q8 How much do you know about running a business?

- A great deal (1)
- A lot (2)
- A moderate amount (3)
- A little (4)
- Nothing at all (5)

Q19 Have you taken this survey before? (Your answer to this question will not affect your pay)

- No (1)
- Yes (2)

Q21 Thank you for all of your work. Please select the names of as many startups as you can remember.

- Teespring (1)
- Appboy (2)
- Lystable (3)
- LendUp (4)
- Homey (5)
- BitGym (6)
- Addepic (7)
- Serefly (8)
- Sackle (9)
- Grativie (10)
- Bandity (11)
- Hacklow (12)
- VomitStain (13)
- CatRash (14)
- Tablebum (15)
- LoanShark (16)
- DoodyCalls (17)
- TortureTour (18)

Q28 What do you think we are going to do with your survey answers?

Q17 Congratulations! You have completed the survey. Here is your survey code:

${e://Field/random_code}

Please enter the code into mturk to be compensated for your work. Thank you!