Optimizing Response Rate in Customer Satisfaction Survey

Datasci 241 - Final Project

Connor Ethan Yen, Matthew Dodd, Leena Bhai, and Heather Rodney

4/19/2023

1 Abstract

Internal IT often seeks to improve customer experience and satisfaction. However, targeting satisfaction through customer satisfaction surveys (CSS) is complicated by the trade off between survey specificity and response rates. Very simple CSS achieve high response rates but provide teams with little actionable direction, while highly detailed CSS risk statistical bias and uncertainty due to low response rates. In this study, we focus on internal IT at the midsize pharmaceutical company and experiment with different types of CSS—ranging from simple thumbs-up/thumbs-down to multiple targeted ordinal questions—to find the type of survey that provides enough actionable and statistically representative information for IT-teams to perform productive work. We find that as surveys increase in length from a baseline of 1 question to 3 and 5 questions, their response rates fell by 17.6% (6.5%) and 34.9% (6.5%) respectively. We further find that there is no statistically significant effect on response rate in using a binary or likert response scale. Finally, we find that there are no heterogenous treatment effects at the departmental or salary grade levels.

2 Introduction

The tradeoff between CSS specificity and response rates is well documented in the literature. Factors such as question length/amount (e.g., Burchell & Marsh 1992, Galesic & Bosnjak 2009, Revilla & Ochoa 2017), survey phrasing (e.g., Ograjenšek & Gal 2011, Nicolini & Valle 2011), and user demographics (e.g., Revilla & Höhne 2020) have all been considered. Nevertheless, little work has been done to quantitatively characterize the tradeoff between CSS specificity and response rates from an experimental point of view.

This study partners with an unspecified midsize pharmaceutical company to experiment on the CSS complexity versus response rate tradeoff. IT teams at this company are interested in improving the quality of service delivered and hope to do so by using customer satisfaction surveys. While complex surveys allow IT teams to parse out individual pain-points, the literature suggests that these surveys also lead to low response rates. On the other hand, simple surveys may achieve high response rates, but lack the specificity required for teams to identify where to improve. This study characterizes survey complexity across two dimensions: number of questions and response scale. We experiment with how response rate is effected by the number of questions asked and the response scale deployed.

2.1 Research Questions and Hypothesis

1. How does increasing the number of questions on a survey effect response rate?
    - We hypothesize that increasing the number of questions on a survey decreases survey response rate.

2. Does changing the response scale from likert to binary increase response rate?
    - We hypothesize that changing the response scale from binary to likert decreases response rate.

3 Experimental Design

When a user raises an issue to the IT help desk, their issue is raised as a "ticket." These tickets document details relevant to the user's case. With the ticket raised, the IT help desk assigns a lead "agent" to work on the issue. Users can check the status of their tickets through a web portal. Once the ticket has been resolved, the user can close the ticket through the portal.

One of the features provided by the service management platform is collection of satisfaction surveys. If enabled, users are directed to a survey page immediately after clicking the "close" button on their ticket. We leveraged this feature to perform our experiment. Our team had the advantage of circumventing prior user bias due to former experiences with the survey process since the survey feature had been disabled in early 2020. As the company nearly doubled in size between 2020 and 2023, most users of the ticket platform had never been exposed to internal help desk surveys before this experiment was run.

The service management platform has a centralized data repository for tickets with rows indexed by unique ticket IDs. Columns for if the survey was sent (True/False), question-1 response (1-5 or +/-), question-2 response (1-5 or +/-), question-3 response (1-5 or +/-), question-4 response (1-5 or +/-), and question-5 response (1-5 or +/-) were added. Thus, survey results populate in the database as field values. We define response to be True if at least one of the values in the "question-1 response" through "question-5 response" columns is not empty. In the case of a positive response, the user successfully clicked "submit" on the survey page. Note that all questions on the survey were required so users who successfully submitted a survey responded to all the survey questions asked. We define response to be False if the "question-1 response" through "question-5 response" columns are all empty. A derived column corresponding to survey response (True/False) was added for ease of analysis.

The column tracking if the survey was sent automatically populates with "False." If upon closure of a ticket, the service management platform directs the users browser to the survey page, the "survey sent" column switches to True. In cases where the user did not respond to the survey (i.e., the survey response column is False), the user either closed out of the survey page or the page timed-out. In cases where the survey redirect was blocked by the user's browser, the "survey sent" column does not always record a consistent value. However, browsers are configured at the enterprise level with default privileges enabling the survey redirect. Even for users interested in blocking redirects from general websites to avoid annoying ads, since the survey link is hosted through the company domain, the survey url is necessarily white-listed. While it is possible for individual users to have configurations that block the explicit survey redirect, this issue should only exist for a small minority of users.

3.1 Treatment

For this study we deployed a 2 by 3 design with dimensions of response granularity and number of questions. Response granularity had two levels: binary (a simple thumbs-up and thumbs-down) and likert (one to five scale). The surveys further had either one, three, or five survey questions. The following questions were used:

1. How satisfied are you with your overall experience of the service desk?
2. How satisfied are you with the resolution speed of your ticket?
3. How satisfied are you with the degree of communication provided during the resolution process?
4. How satisfied are you with the usability of the platform?
5. How satisfied are you with the resolution provided by the service desk?

The syntactic question format of "how satisfied are you with _____?" was chosen intentionally to keep the complexity of each question constant. Individuals in the study thus receive one of six treatments:

1. Question 1 with binary scale
2. Question 1 with likert scale
3. Questions 1-3 with binary scale
4. Questions 1-3 with likert scale
5. Questions 1-5 with binary scale
6. Questions 1-5 with likert scale

Note that binary ratings presented as a thumbs-up or thumbs-down scale under each question while the likert ratings presented as a 1 through 5 scale with 1 corresponding to "not at all satisfied" and 5 corresponding to "very satisfied."

3.2 Enrollment and Randomization

Surveys were only enabled for each user once. That is, for each user (defined by Active Directory profiles), only the first ticket "closed" during the experimental timeframe received a survey redirect. Randomization was conducted at ticket closure. The process flow is as follows:

1. The user clicks "close" on their ticket.
2. The platform database collates the user ID on tickets where a survey was sent (i.e. "sent survey" == True).
3. If the user ID associated with the ticket is NOT in the collated list from step-2, the user is randomly redirected to one of the six surveys (i.e., treatments).
4. If the user ID associated with the ticket is in the collated list from step-1, the user is redirected to their profile home page.

This method ensures that each "participant" of the study is unique and avoids dependence between data points. The power analysis for this study demonstrated that to observe a large true effect (~10% difference between treatments), our sample size should be over 500, and to observe a small true effect (~3% difference between treatments), our sample size should be over 2500. Our study involved 672 individuals providing us the power to observe a large true effect.

3.3 Experimental Timeline

The experiment was conducted over a 3 week period from February 13, 2023 to March 3, 2023 (PST). This period was chosen to not correspond with any major business processes (e.g., financial quarter ends) that may have uniformly influenced the urgency of work and subsequent inclination to engage in non-critical tasks.

The confined time frame also reduced spillage between treatment groups since as the experimental window increases, discussion between coworkers might cause individuals to realize that different satisfaction surveys were being conducted. Nevertheless, because survey response is not a metric reported at an executive level through which survey response would be heavily encouraged, we do not expect spillover effects to be significant.

3.4 Data Completeness and Omitted Covariate Bias

Noncompliance occurs when a user intentionally circumvents the survey redirect after clicking "close" on their ticket. In these cases, we intend for the users to be treated with a survey, but treatment is not delivered.

This can occur if the user's enables their browser to block the survey link. However, as mentioned earlier, these settings only exist for a small subset of users. Since $CACE = ITT/\alpha$, for small non-compliance, $\alpha \to 1$, and $CACE \approx ITT$.

Several key covariates have been omitted due to privacy concerns expressed by the partner company. Potential covariates that might effect response rate are demographic covariates such as age, race, and gender. Moreover, ticket-related covariates such as ticket type and priority plausibly effect response propensity.

4 Exploratory Data Analysis

4.1 Average Treatment Effect

We first tabulate the ATE for each treatment (see Table 1) along with the treatment population size.

Table 1: Response rate across different survey types.

Number of Questions	Response Scale	Count	Survey Respondees	Response Rate
1 question	binary	107	68	63.6%
1 question	likert	116	67	57.8%
3 questions	binary	112	51	45.5%
3 questions	likert	121	51	42.1%
5 questions	binary	113	32	28.3%
5 questions	likert	103	24	23.3%

Our analysis reveals a negative correlation between survey complexity and response rate. The simplest survey type, featuring a single binary question, yields the highest response rate at 63.5%. In contrast, the most intricate survey, consisting of a 5-item Likert scale questionnaire, demonstrates the lowest response rate of 23.3%. This suggests that as the survey's complexity increases, participant engagement tends to decrease. Interestingly, when the number of questions is held constant, changing the response scale from binary to likert only marginally changes response rate. These preliminary results that we are observing a large treatment effect with number of questions but a small treatment effect with response scale.

4.2 Covariates

The data has columns for department, salary grade, ticket resolution time, the number of questions on the survey, the survey response scale, and if the user responded to the survey. Department, salary grade, and ticket resolution time represent covariates. It is possible that different cultures associated with different departments and salary grades influence CSS response propensity. Moreover, as the time taken to resolve the ticket increases, we generally expect users to feel more unhappy—it is likely that this affective component also influences response rate. For example, we expect the response rate for junior employees in IT with short ticket times to be different than the response rate for senior employees in Finance with long ticket times. Due to privacy concerns from the partner company, the values for the department and salary grade covarites have been anonymized. We first examine the salary grade and departmental distributions in tables 2 and 3.

Table 2: Distribution of salary grades amongst experiment participants

Grade	Count
grade 1	42
grade 2	124
grade 3	506

Table 3: Distribution of departments amongst experiment participants

Department	Count
A	139
B	212
C	321

There are substantially more individuals in salary grade 3 (506) than in grade 1 (42). This demonstrates that the covariate for salary grade is not uniformly distributed within our experiment. We now examine the distribution of the salary grade and department covarites amongst our treatments to check for covariate imbalance in table 4.

Table 4: Covariate distribution amongst treatments

Questions	Scale	Salary Grade A	Salary Grade B	Salary Grade C	Dpt A	Dpt B	Dpt C
1 question	binary	5	20	82	14	40	53
1 question	likert	7	23	86	30	37	49
3 questions	binary	10	14	88	22	29	61
3 questions	likert	8	29	84	29	41	51
5 questions	binary	4	23	86	23	36	54
5 questions	likert	8	15	80	21	29	53

It appears that the distribution of covariates across treatments are similar—we will formally test for this in the next section. We further examine the distribution of ticket response times in Figure 1.

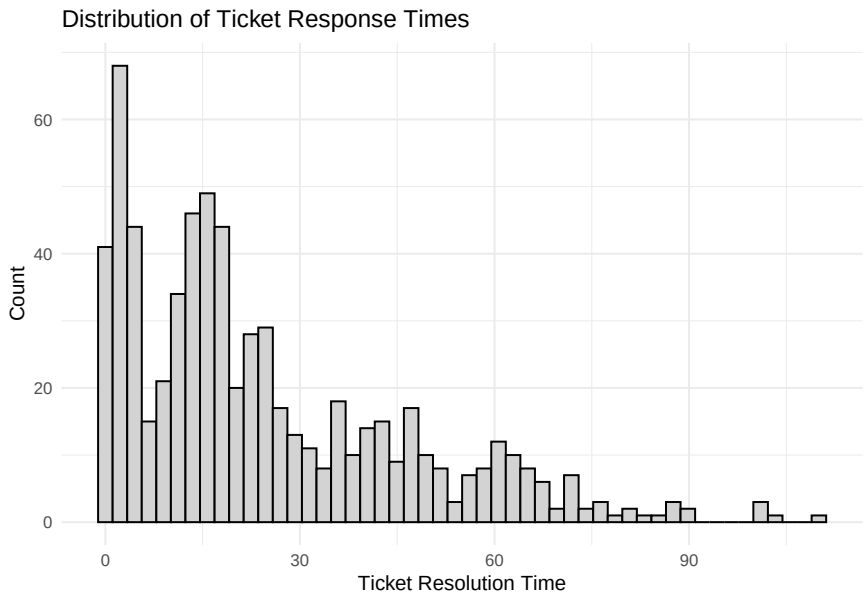

Figure 1: Histogram of ticket response times for experiment participants.

The histogram exhibits a right-skewed distribution, with most tickets being resolved within a day, while some tickets taking about three days to resolve. We therefore apply a logarithmic transformation to resolution time when including it as a covariate.

4.3 Check Random Treatment Assignment

To ensure that treatment assignment is random with respect to the covariates, we can compare two linear models for treatment assignment where the first model omits covariate terms and the second model includes them. Because we have six treatments in total, this comparison has to be performed six times where treatment assignment is coded as an indicator variable for receiving a particular treatment. The first model is of the form:

$$treatment_i = \beta_0,$$

where β_0 is the average rate of treatment, and $treatment_i$ is an indicator variable for one of the six treatment groups. The second model is of the form:

$$\begin{aligned} treatment_i = \beta_0 &+ \beta_1 \mathbf{1}_{dpt=B} + \beta_2 \mathbf{1}_{dpt=C} \\ &+ \beta_3 \mathbf{1}_{grd=2} + \beta_4 \mathbf{1}_{grd=3} \\ &+ \beta_5 \mathbf{1}_{dpt=A}\mathbf{1}_{grd=1} + \beta_6 \mathbf{1}_{dpt=A}\mathbf{1}_{grd=2} + \beta_7 \mathbf{1}_{dpt=A}\mathbf{1}_{grd=3} \\ &+ \beta_8 \mathbf{1}_{dpt=B}\mathbf{1}_{grd=1} + \beta_9 \mathbf{1}_{dpt=B}\mathbf{1}_{grd=2} + \beta_{10} \mathbf{1}_{dpt=B}\mathbf{1}_{grd=3} \\ &+ \beta_{11} \mathbf{1}_{dpt=C}\mathbf{1}_{grd=1} + \beta_{12} \mathbf{1}_{dpt=C}\mathbf{1}_{grd=2} + \beta_{13} \mathbf{1}_{dpt=C}\mathbf{1}_{grd=3} \\ &+ \beta_{14} \log(res), \end{aligned}$$

where β_i for represent linear regression coefficients and $\mathbf{1}_*$ represent indicator variables that take the value 1 when the subscript condition is true. "Department," "salary grade," and "ticket resolution time" have been abbreviated as "dpt," "grd," and "res" respectively. Simply put, the second model regresses the treatment group on all the combination sof the covariates to see if being in a particular department and salary grade causes the individual to be in a particular group. We then compare the two models with an F-test and report p-values in the table below.

Table 5: Test if treatment assignment is random with respect to covariates

treatment	Pr(>F)
1 question + binary	0.3838967
1 question + likert	0.8325269
3 questions + binary	0.0738914
3 questions + likert	0.1269439
5 questions + binary	0.6129196
5 questions + likert	0.7037569

None of the p-values in table 5 are significant at a 5% level. Thus, covariates (and combinations therein) for department, salary grade, and resolution time does not explain treatment assignment better than taking a simple average rate of treatment.

5 Results

5.1 Regression

In order to analyze the data we specified three different regression models. Our initial baseline model regresses our binary "responded to survey" variable on the number of questions contained in the survey which took the value of 1, 3 or 5. We found that surveys with one question had a response rate of 60.5%.

The regression showed that as surveys increased in length from 3 to 5 questions their response rates fell by -17 and -34.6% respectively, both statistically significant results.

Our second specification added the binary/likert response scale treatment. Interestingly, this specification shows that there is no statistically significant effect of response scale to response rates for any of the survey lengths even though there is a strict decrease in response rate between binary and likert scaled surveys when holding survey length constant.

Our final model adds to the interaction specification by adding controls for respondant department, salary grade and the log of a tickets response time. This log transform was used to the the right hand skew in our ticket resolution time series. The results again show that it only seems that survey length has a statistically significant effect on response rates, however the added controls do increase our point estimates for the decrease in response rates for 3 and 5 questions to -17.6% (6.5%) and -34.9% (6.5%) respectively. This final model is of the form:

$$\begin{aligned} R = \beta_0 &+ \beta_1 \mathbf{1}_{q=3} + \beta_2 \mathbf{1}_{q=5} + \beta_3 \mathbf{1}_{likert} \\ &+ \beta_4 \mathbf{1}_{dpt=B} + \beta_5 \mathbf{1}_{dpt=C} + \beta_6 \mathbf{1}_{grd=2} + \beta_7 \mathbf{1}_{grd=3} + \beta_8 \log(res) \\ &+ \beta_9 \mathbf{1}_{q=3} \mathbf{1}_{likert} + \beta_{10} \mathbf{1}_{q=5} \mathbf{1}_{likert}, \end{aligned}$$

where R is response rate, subscript q refers to number of questions, and "department," "salary grade," and "ticket resolution time" have been abbreviated as before. Results are tabulated in the table below.

Table 1: Comparison Between Regression Coefficients in Causal Linear Models

	Dependent variable:		
	(1)	(2)	(3)
3 Questions	-0.170^{***}	-0.170^{***}	-0.176^{***}
	(0.045)	(0.045)	(0.065)
5 Questions	-0.346^{***}	-0.348^{***}	-0.349^{***}
	(0.046)	(0.046)	(0.065)
Likert		-0.049	-0.056
		(0.037)	(0.064)
Salary Grade 2			-0.076
			(0.086)
Salary Grade 3			-0.117
			(0.078)
Department B			0.067
			(0.052)
Department C			0.024
			(0.049)
log(Resolution Time)			-0.004
			(0.014)
3 Questions + Likert			0.013
			(0.090)
5 Questions + Likert			0.005
			(0.092)
Constant	0.605^{***}	0.631^{***}	0.713^{***}
	(0.032)	(0.037)	(0.104)
Note:			*p<0.1; **p<0.05; ***p<0.01

5.2 Heterogenous Treatment Effects

We test for heterogeneous treatment effects by interacting the treatment variables with different covariates. Results are summarized in the table below.

Table 2: Regression for Heterogenous Treatment Effects

	Dependent variable:
3 Questions	-0.545^{***} (0.186)
5 Questions	-0.769^{***} (0.217)
Likert	-0.153 (0.103)
Dpt B	-0.240 (0.194)
Dpt C	-0.376^{*} (0.214)
Grd 2	-0.401^{**} (0.190)
Grd 3	-0.524^{***} (0.150)
log(res)	-0.006 (0.015)
3 Questions + Likert	0.041 (0.097)
5 Questions + Likert	0.028 (0.093)
Dpt B + Grd 2	0.134 (0.227)
Dpt C + Grd 2	0.248 (0.236)
Dpt B + Grd 3	0.182 (0.191)
Dpt C + Grd 3	0.300 (0.206)
3 Questions + Dpt B	0.175 (0.134)
5 Questions + Dpt B	0.102 (0.131)
3 Questions + Dpt C	0.126 (0.125)
5 Question + Dpt C	0.077 (0.120)
3 Questions + Grd 2	0.145 (0.194)
5 Questions + Grd 2	0.367^{*} (0.216)
3 Questions + Grd 3	0.278 (0.169)
5 Questions + Grd 3	0.356^{*} (0.190)
Likert + Dpt B	0.096 (0.108)
Likert + Dpt C	0.111 (0.099)
Constant	1.205^{***} (0.164)
Note:	*p<0.1; **p<0.05; ***p<0.01

When the treatment variables interact with covariates for department and salary grade, we see that the coefficients for both salary grades 2 and 3 are significant. However, none of the interaction terms between treatment and salary grade are significant at a 95% level. Thus, we conclude that there is no observable heterogeneous treatment effect for the covariates included in this study.

5.3 Generalizability and Limitations

The experimental population was defined as the set of individuals who "closed" IT Help Desk Tickets between February 13, 2023 and March 3, 2023 (PST), where treatment is only assigned the *first time* an individual "closes" a ticket during that time period. Over Q1 2023, the partner company saw rapid growth following several key scientific and business successes. As such, during the duration of the study, there was a large inflow of new employees and contractors which enabled us to collect more unique data points than initially anticipated. However, this study only examines response rate the *first time* an individual receives a survey. As the company's head count stabilizes, repeated survey results will be pulled from a relatively constant population. This study does not quantify the survey complexity versus response rate tradeoff when surveys

are requested multiple times to the same population which is frequently the use-case for teams seeking to build satisfaction over time metrics.

Nevertheless, the results from this study have are useful for teams that agile frameworks. For example, if teams use customer satisfaction surveys to identify or verify actionable pain-points for a discrete subset of individuals, these surveys only need to be sent to those smaller set of individuals. After teams have received results from the survey and worked to improve service accordingly, the next step would be to identify pain-points for a *new* subset of individuals. This process of iterating over targeted user subgroups means that when teams return to a subgroup that had received a survey in the past, a substantial amount of time would have elapsed. This time difference between surveys allows response propensity behavior on the second survey to return to that for the novel first survey. Note that agile teams frequently do not benefit from surveying aggregate populations as discrete action items are hard to identify for very large groups. Thus, this method of iterative surveying over small user subgroups aligns well with agile team workflows.

6 Conclusion

We experiment to quantify how increasing survey complexity decreases response rate. We find that increasing the number of questions on a survey from 1 to 3 questions and then from 1 to 5 questions results in a -17.6% (6.5%) and -34.9% (6.5%) decrease in response rate, respectively. Interestingly, we find that changing the response scale from binary to likert, while keeping the number of questions constant, did not show a statistically significant decrease in response rate. This is not to say, however, that no effect exists. Recall that our power analysis found that to observe a small effect (of the magnitude ~3%), we would need over 2500 individuals. It is likely the case that the treatment effect of changing the response scales on surveys is to small for our experiment to detect. Also of note is that while the first-time survey response rate for the simplest survey (binary 1 question) is high, at 63.6%, further surveys on the same population are likely to experience rapidly declining response rates. Future work should be done to measure the time component of survey treatment. Nevertheless, this study is valuable in showing how the survey complexity to response rate trade off can be quantified in ways that allow teams to design the ideal survey that gives enough information while also maintaining adequate sample representation.

7 References

1. Burchell, Brendan, and Catherine Marsh. "The effect of questionnaire length on survey response." *Qual Quant*, Vol. 26, 1992, pp. 233–244. https://doi.org/10.1007/BF00172427.
2. Galesic, Mirta, and Michael Bosnjak, "Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey," *Public Opinion Quarterly*, Vol. 73, Issue 2, Summer 2009, pp. 349–360, https://doi.org/10.1093/poq/nfp031.
3. Nicolini, Giovanna, and Luciana Dalla Valle. "Census and Sample Surveys." Modern Analysis of Customer Surveys, *John Wiley & Sons, Ltd*, 2011, pp. 37–53, https://doi.org/10.1002/9781119961154.ch3.
4. Ograjenšek, Irena, and Iddo Gal. "The Concept and Assessment of Customer Satisfaction." Modern Analysis of Customer Surveys, *John Wiley & Sons, Ltd*, 2011, pp. 107–27, https://doi.org/10.1002/9781119961154.ch7.
5. Revilla, Melanie, and Carlos Ochoa. "Ideal and Maximum Length for a Web Survey." *International Journal of Market Research*, Vol. 59, Issue 5, 2017, pp. 557–565. https://doi.org/10.2501/IJMR-2017-039.
6. Revilla, Melanie, and Jan Karem Höhne. "How long do respondents think online surveys should be? New evidence from two online panels in Germany." *International Journal of Market Research*, Vol. 62, Issue 5, 2020, pp. 538–545. https://doi.org/10.1177/1470785320943049.