Preventing Satisficing in Online Surveys:

A "Kapcha" to Ensure Higher Quality Data *

Adam Kapelner†
The Wharton School
Department of Statistics
3730 Walnut Street
Philadelphia, PA, 19104
kapelner@wharton.upenn.edu

Dana Chandler‡
Massachusetts Institute of Technology
Department of Economics
50 Memorial Drive
Cambridge, MA, 02142
dchandler@mit.edu

ABSTRACT

Researchers are increasingly using online labor markets such as Amazon's Mechanical Turk (MTurk) as a source of inexpensive data. One of the most popular tasks is answering surveys. However, without adequate controls, researchers should be concerned that respondents may fill out surveys haphazardly in the unsupervised environment of the Internet. Social scientists refer to mental shortcuts that people take as "satisficing" and this concept has been applied to how respondents take surveys. We examine the prevalence of survey satisficing on MTurk. We present a question-presentation method, called *Kapcha*, which we believe reduces satisficing, thereby improving the quality of survey results. We also present an open-source platform for further survey experimentation on MTurk.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sciences; H.5.m [**Informational Interfaces and Presentation (e.g., HCI)**]: Information Systems

General Terms

Economics, Experimentation, Human Factors, Design, Measurement

Keywords

Mechanical Turk, crowdsourcing, survey design, online surveys, satisficing, instructions, screening, manipulation checks

1. INTRODUCTION

It has been well established that survey-takers may "satisfice" (i.e., take mental shortcuts) to economize on the amount of effort and attention they devote to filling out a survey [12].[1] As a result, the quality of data in surveys may be lower than researchers' expectations. Because surveys attempt to measure internal mental processes, they are by their very nature not easily verifiable by external sources. This presents a potential problem for the many researchers who are beginning to employ Amazon's Mechanical Turk (MTurk) workers to answer surveys and participate in academic research[17, 10, 4]. Moreover, unlike other tasks completed on MTurk, inaccuracies in survey data cannot be remedied by having multiple workers complete a survey, nor is there an easy way to check them against "gold-standard" data.[2]

In our experiment, we examine alternative ways to present survey questions in order to make respondents read and answer questions more carefully.

Our first treatment "exhorts" participants to take our survey seriously. We ask for their careful consideration of our questions by placing a message in prominent red text on the bottom of every question. Surprisingly, this has no effect.

Our two other treatments took a more economic approach and attempted to alter the incentives of survey-takers who ordinarily have an incentive to fill out questions as quickly as possible in order to maximize their hourly wage and exert minimal cognitive effort. More specifically, both treatments force the participant to see the question for a certain "waiting period". Combined, these waiting period treatments improved the quality of survey responses by 10.0% ($p < 0.001$). Under the waiting period treatments, the participant is forced to spend more time on each question and once there, we hypothesize that they will spend more time thinking about and thoughtfully answering questions.

The first of these two treatments, called simply the *Timing control* treatment, features a disabled continue button for the duration of the waiting period. The second of these treatments, referred to as the *Kapcha*[3] has a waiting period equal to that of the *Timing control* treatment, but also attempts to attract the attention of respondents by sequen-

*This paper is a collaborative effort where each author contributed equally.

†Principal corresponding author

‡Corresponding author.

CrowdConf 2010, October 4, 2010, San Francisco, CA.

[1]Jon Krosnick [12] applies Herbert Simon's famous idea of satisficing [18] to how respondents complete surveys.

[2]In an MTurk context, "gold-standard" data refers to asking workers questions to which the surveyor already knows the answer as a way to identify bad workers. Although this is straightforward for an image labeling task (e.g. [9] and [19]), it is less clear how to apply this concept to surveys.

[3]The name was inspired from the "Captcha" Internet challenge-response test to ensure a human response [23].

tially "fading in" each word in the question's directions, its prompt, and its answer choices. This treatment was the most effective and improved quality by approximately 13%.

To proxy for quality, which is largely unobservable, we introduce a "trick question" into the survey as a way of measuring whether people carefully read instructions. We echo the methodology from [16] who call this trick question an *instructional manipulation check* (IMC). Additionally, we give respondents two hypothetical thought experiments where we ask them to imagine how they would behave under certain conditions. The conditions are identical except for a *subtle word change* that would only be apparent if the instructions were read carefully — hence, for close readers, there should be a greater difference in reported behavior as compared with people who were merely skimming.[4]

This paper presents initial evidence on alternative ways to present survey questions in order to reduce satisficing. We hypothesize that altering the cost-benefit analysis undertaken by survey respondents is the mechanism which reduces satisficing. The approach we present has the benefit of improving the quality of results without increasing monetary cost or convenience for the surveyor. We also examine the prevalence of satisficing and how it may vary across respondent demographics. Finally, we discuss ideas for further improving how to present survey questions.

Section 2 explains our experimental methods. Section 3 illustrates our most important results, section 4 concludes, and section 5 talks about future directions. Appendix A describes the TurkSurveyor open-source package for running experiments and appendix B provides links to our source code and data so that others may replicate and verify our results.

2. METHODS

2.1 Recruitment of Participants

We designed an MTurk HIT (our "task") to appear as a nondescript survey task similar to many others that are now popular on MTurk. By making our survey appear like any other, we intended to recruit a population that is representative of the MTurk survey-taking population.

We entitled our task "Take a short 30-question survey — $0.11USD" and its description was "Answer a few short questions for a survey". The preview pane of the HIT only displayed, "Welcome to the short survey! In this survey you will answer 30 questions. You may only do one survey." Our HIT was labeled with the keywords "survey", "questionnaire", "survey, "poll", "opinion", "study", and "experiment" so that people specifically looking for survey tasks could easily find our task. However, we also wanted to attract workers who were not specifically looking for survey tasks. Therefore, we posted batches of tasks at various times.

We recruited 784 MTurk workers from the United States.[5] Each worker was only allowed to complete one survey task and took one of four different versions of the survey according to their randomized assignment. In total, 727 workers completed the entire survey and each was paid $0.11USD.[6]

We posted in batches of 200 tasks at a time four times per day (at 10AM, 4PM, 10PM, and 4AM EST) as to not bias for early-riser or night-owl workers. Altogether, 18 bunches (of 200 HITs each) were posted between September 1 and September 5, 2010. All HITs expired six hours after creation as to not interfere with the subsequent batch. Note that if we had posted our tasks as one gargantuan batch and waited until completion (possibly a week or longer), we would have attracted a majority of workers who were *specifically* looking for survey tasks (most likely searching for them via keyword) rather than a more general sample of workers.[7] The workers were given a maximum of 45 minutes to complete the task.[8]

2.2 Treatments

To test the satisficing-reducing effect of the *Kapcha*, we randomized each participant into one of four treatments which are described below, summarized in table 1, and pictured in figures 1a–d.[9]

Table 1: Overview of treatments and how they improve data quality

Treatment	Increase perceived value of survey	Force slow down	Attract attention to individual words
Control			
Exhortation	✓		
Timing control		✓	
Kapcha		✓	✓

The first treatment was the *Control* where the questions were displayed in a similar fashion as any other online survey.

Our *Exhortation* treatment presents survey questions in an identical way as the *Control* treatment except that we try to increase the survey taker's motivation by reminding them in alarming red text at the bottom of each question page to "Please answer accurately. Your responses will be used for research." Past research on survey design has shown that respondents are more likely to devote effort to complet-

[4]This portion of our experiment replicates Study 1 in [16]. Their primary focus was to identify subsamples of higher quality data and to eliminate the "noisy data" (i.e., the participants who did not read the instructions carefully enough to pass the trick question). This enables researchers to increase the statistical power of their experiments.

[5]This restriction reduces the influence of any confounding language-specific or cultural effects. Generalizing our results to other countries and languages may be a fruitful future research direction.

[6]The workers worked 81.3 hours at an average wage of $0.98/hr and a total cash cost to the experimenters of $87.97 (including Amazon's fee of 10%). In this calculation, we ignore the time they spent on the feedback question and the bonuses we paid.

[7]This phenomenon is due to HITs rapidly losing prominence in the public listings and eventually being relegated to obscurity where they may only be found by those searching via keyword (see [5]). Also note that we save the time each HIT was created at and expires at. We use this information to check at what point in the HIT listing life-cycle the worker accepted the HIT.

[8]Even though the survey was short, we wanted to give ample time to be able to collect data on task breaks. Note that few workers took advantage of the long time limit; 117 workers (16%) took more than 15 minutes and only 61 workers (8%) took more than 30 minutes.

[9]Videos llustrating the four treatments are available at http://danachandler.com/kapchastudy.html

ing surveys if they perceive them as valuable because they contribute to research [12].

Rather than using exhortation, our *Timing control* and *Kapcha* treatments induce more careful survey taking by changing the incentives of a respondent. In short, we lower the payoff to satisficing.[10] When respondents can breeze through a survey and click one answer after another without delay, they may be tempted to satisfice — i.e., click the first answer that seems correct or any answer at random. If, however, survey respondents are forced to wait before proceeding to the next question, we hypothesize that they will use this time to think more carefully about how to answer.

Our *Timing control* treatment is identical to the *Control* treatment except that the continue button is disabled and has a spinning graphic during a waiting period[11] after which the continue button is enabled.

The *Kapcha* treatment goes one step further and, in addition to slowing down the respondent for a time equal to the *Timing control* treatment, also draws additional attention to the instructions and answer choices by "fading-in" the survey's words at 250 words per minute.

The delay time for the questions in the *Timing control* treatment were calibrated to be the same total fade-in time for the *Kapcha* participant's question. By controlling for the timed delay, we were able to isolate the additional effect due to forcing the respondent to pay attention to the words in the *Kapcha*.

Although this is the first research to our knowledge that studies waiting periods and textual fade-ins, there is a long history of research on how various forms of survey implementation affect response. Two interesting examples include how self-administration lead respondents to answer sensitive questions more truthfully and how questions that are accompanied by audio do the same among people who might not understand the text (especially among low-literacy respondents). Recently, [6] has helped separate the effect of the self-administration and the audio component.[12]

2.3 Custom Survey Task Design

As soon as the worker accepted the HIT, they were given a page with directions that explained the length of the survey and asked to begin when ready. Depending on the treatment, we also added an additional sentence or two to the instructions in order to explain the particularities associated with each treatment. For our *Exhortation* group, we emphasized the importance of giving accurate and honest answers. In our *Timing control* group, we told participants that the continue button would be disabled for a short time so they would have more time to read and answer each question. For our *Kapcha* group, we mentioned how words and answer choices would appear one at a time.

After reading the directions, the worker began the survey task which consisted of 30 questions plus two optional questions eliciting feedback. Each question was presented individually so that the respondent must click submit before moving onto the next question.[13]

Our first question, "question A", is a hypothetical thought experiment (which we call the soda-pricing example) that "demonstrates how different expectations can change people's willingness to pay for identical experiences" [16]. The question text is shown below. The subtle text manipulation which induces an effect according to [21] is shown in brackets and will be denoted as the "run-down" vs. "fancy" treatments:

> You are on the beach on a hot day. For the last hour you have been thinking about how much you would enjoy an ice cold can of soda. Your companion needs to make a phone call and offers to bring back a soda from the only nearby place where drinks are sold, which happens to be a [run-down grocery store / fancy resort]. Your companion asks how much you are willing to pay for the soda and will only buy it if it is below the price you state. How much are you willing to pay?

It has been shown repeatedly in the literature that people are willing to pay more when the beverage comes from a fancy resort. If the workers were reading the instructions carefully, we expect them to pay a higher price in the "fancy" treatment.

The worker was then given "question B," another hypothetical thought experiment (which we call the football attendance example), which demonstrates that people are susceptible to the sunk cost fallacy (for screenshots, see figure 1a–d). The question text is shown below. The subtle text manipulation which induces an effect according to [21] is shown in brackets and will be denoted as the "paid" vs "free" treatments:

> Imagine that your favorite football team is playing an important game. You have a ticket to the game that you have [paid handsomely for / received for free from a friend]. However, on the day of the game, it happens to be freezing cold. What do you do?

Their intention was gauged on a nine-point scale where 1 was labeled "definitely stay at home" and 9 was labeled "definitely go to the game". It has been shown that people who read the treatment where they paid for the tickets are more likely to go to the game.

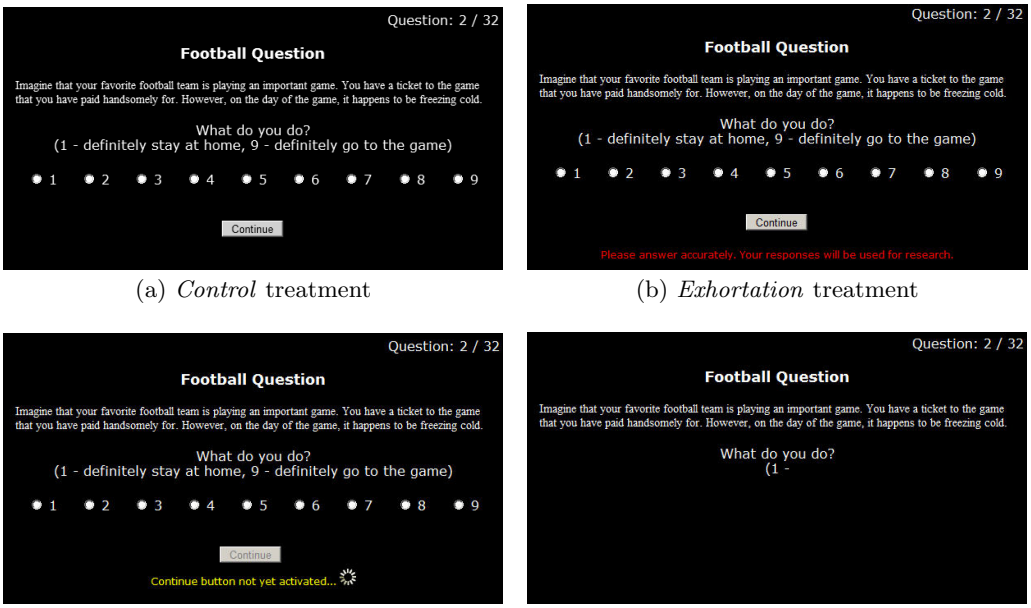
By randomizing the text changes independently of treatments, we were able compare the *strength* of these two well-established psychological effects across the four types of survey-presentation.

[10] If the *Exhortation* treatment increased the rate at which people passed the trick question (which it did not), we might have worried that this framing could bias the way survey takers answer questions, particularly socially sensitive ones, since it reminds the respondent that they are under scrutiny. In social psychology, over-reporting "positive" behaviors is known as the "social desirability bias" [7].

[11] We peg the waiting period to the time it takes an average person to read the number of words in each question. [20] finds that the average reading speed for college-level readers is 280 words per minute and 250 for twelfth-graders. We chose 250 words per minute.

[12] For an excellent, though slightly dated review of various survey presentation formats and the issues they try to overcome, see Chapter 10.1 of [22]

[13] Note that most surveys on MTurk display all questions on one page. Presenting questions one at a time, as we do in our study, probably serves to reduce satisficing. A future study would allow us to determine how the number of questions on each page affects satisficing.



(a) *Control* treatment

(b) *Exhortation* treatment

(c) *Timing control* treatment

(d) *Kapcha* treatment

Figure 1: The participant's screen during question B, the intent to go to a football game, shown for all four experimental treatments (in the "paid" treatment). The *Timing control* and *Kapcha* treatments are both shown at 10 seconds since page load.

The participant then answered an "instructional manipulation check" (IMC) question. Once again, this is a trick question designed to gauge whether the participant reads and follows directions carefully. The instructions of the IMC question asks the participant which sports they like to play and provides many options. However, within the question's directions (i.e. the "fine print"), we tell the respondent to ignore the question prompt and instead click on the title above (see Fig 2). If the participant clicked the title, they "pass" the IMC; any other behavior is considered failure. We administer the IMC and consider it to be a proxy for satisficing behavior in general[14] which we were able to compare across the four treatments.

After this point, there are no longer any differences in how we present survey questions across the treatments. We turn off the *Kapcha* fade-in, the *Timing control*, and the *Exhortation* message. This allows us to collect demographics and need for cognition measures in a way that is comparable and independent of treatment.[15]

The next eight questions collect demographic information. We ask for birth year, gender, and level of education. We then ask a few questions about their general work habits on MTurk: "Why do you complete tasks in Mechanical Turk?", "How much do you earn per week on Mechanical Turk?", "How much time do you spend per week on Mechanical Turk?"[16], and "Do you generally multi-task while doing HITs?" To see if people who take surveys more often are any different, we also ask: "What percent of your time on MTurk do you spend answering surveys, polls, or questionnaires?".

After the eight demographic questions, we administer an 18-question abbreviated version of the full "Need for Cognition" (NFC) scale (see [2]). Respondents say how characteristic each of the statements are of their personality (e.g. "I find satisfaction in deliberating hard and for long hours") on a five-point scale where 1 indicates "extremely uncharacteristic" and 5 indicates "extremely characteristic".[17] In short, the NFC scale assesses how much an individual has a need to think meticulously and abstractly. This is of interest to our study because an individual's NFC may itself affect the probability of satisficing during a survey. People who have a higher NFC are also more likely to fill out the survey diligently and appear with high scores. Those with low NFC will have scores that are attenuated toward the center (i.e., 3) because they are more likely to haphazardly guess.

For the 30th question, we ask the participant to rate how motivated they were to take this survey on a nine-point scale.

We then give two *optional* feedback questions. On both questions, we indicated that responses would be given either a \$0.01, \$0.05 USD bonus, or no bonus (these three levels were randomized in order to test the effect of bonus level on feedback quality). The amount of feedback is also used as a

[14] [16] could not detect a significant difference between the "fancy resort" and the "run-down grocery store" conditions in question A using the full sample, *but* could detect a significant difference using data only from the participants who passed the IMC. This is an intuitive result; the psychological effect that occurs due to subtle word changes would only be detectable if a respondent carefully read the instructions.

[15] Naturally, the influence of the treatment from the first three questions may linger.

[16] This question was used in [11], which presents result from a large MTurk survey

[17] Approximately half of the questions are also reverse-coded so that noisy survey responses would cancel themselves out and tend toward the center.

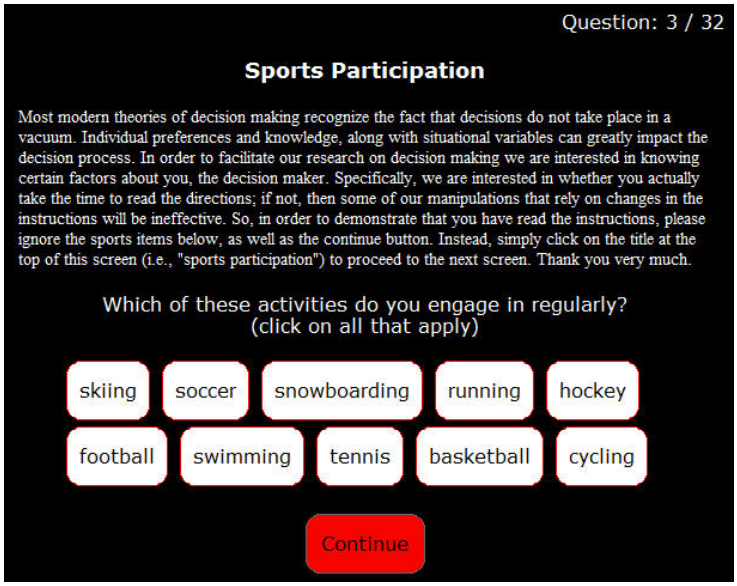


Figure 2: A screenshot of the *instructional manipulation check* in the *Control* treatment.

measure of respondent engagement.

The first question inquired, "What did you like most about this survey? What did you like least about this survey? Is there anything you would recommend to make it better?" The second feedback prompt was only relevant for the *Kapcha* or *Timing control* treatments. We asked, "Certain respondents had to wait for each survey question to complete before filling in answers. We are especially interested in knowing how this affected the way you took the survey."

2.4 Other Data Collected

During our recruitment period, we posted HIT bunches (see section 2.1 for details) with an equal number of tasks in each of the four experimental treatments.

In addition to the participants' responses, we recorded how long survey respondents spent on each part of the survey. This gave us some indication of how seriously people took our survey (i.e., read instructions, considered answers to questions). We also recorded when the participant's task window was focused on our task or focused on another window.[18]

In the future, we hope to collect much more detailed information on the user's activity including timestamps of exact mouse position locations, mouse clicks, and keystrokes. Ultimately, it would be an asset to researchers to be able to "playback" the task by watching the worker's mouse movements in a short video in order to gain greater insight into how respondents answer surveys.[19] This would help identify satisficing behavior in a way that would go undetected using other rigid rules, but would be obvious from watching a video (e.g., instantaneously and haphazardly clicking on random answer choices).

3. RESULTS

Table 2 shows the main results of the experiment. In short, we find that the *Kapcha* treatment increases the proportion of respondents that pass the instructional manipulation check (the *IMC pass rate*) relative to other treatments but causes more people to leave the task midway. We find the *Kapcha* treatment induces a highly significant effect on question A, but not on question B. Overall, we confirm both of Thaler's [21] economic effects if we combine all treatments.

Table 2: Summary statistics by treatment

	Overall	*Control*	*Exhortation*	*Timing*	*Kapcha*
N	784	178	208	210	188
Attrition (%)	7.3	5.6	2.4	8.1	13.3
IMC Pass Rate (%)	81.7	77.4	76.4	83.9	90.2
Question A price ($)					
"fancy"	2.21	2.14	2.17	2.27	2.23
"rundown"	1.97	2.06	2.10	1.96	1.69
difference	0.24**	0.08	0.07	0.31	0.54***
Question B intent					
"paid"	7.28	7.55	7.31	7.14	7.16
"free"	6.91	6.79	7.24	6.35	7.12
difference	0.37*	0.76	0.07	0.79*	0.04

*p < 0.05, **p < 0.01, ***p < 0.001

3.1 Timed Treatments Lead to More Attrition

Table 3 shows the observed attrition for each treatment as well as comparisons against the *Control*.[20] Our two timed

[18]Unfortunately, this variable was not compatible with all Internet browsers and was too noisy to use in analysis.

[19]Everything mentioned here is possible by using http://clicktale.com's premium service.

[20]We employ two-sample two-sided z-tests for difference in proportion.

treatments *timing control* and *Kapcha* led to higher attrition as compared with the *Control* treatment. This result is not surprising since by forcing some respondents to spend more time on our survey, we effectively are lowering their hourly wage and testing their patience. As we might expect, the *Exhortation* treatment, which reminds the respondent of the importance of our survey, slightly lowers attrition (though not significantly).[21]

Table 3: Attrition by treatment

Treatment	N	Attrition (%)	comparison with *Control* treatment (p-value)
Control	178	5.6	—
Exhortation	208	2.4	0.12
Timing control	210	8.1	0.46
Kapcha	188	13.3	0.02
All	784	7.3	—

Ordinarily, survey designers seek to minimize attrition (i.e., maximize completion rates of their surveys). However, in the MTurk environment, where the number of potential respondents is larger than the desired sample size, the researcher may want to restrict the sample to those who yield the highest quality data.[22] If the decision to leave a survey midway through indicates that these people are "less serious", we probably would not want them in our sample.[23]

Note that going forward, we only analyze tasks that were fully completed (i.e. workers who did not attrit).

3.2 Kapcha Alone is a Successful Mechanism for Reducing Satisficing

We investigate the IMC pass rate by experimental treatment and demographic controls.

Table 4: IMC pass rate (%) by treatments with comparisons

Treatment	IMC Pass Rate (%)	comparisons (p-value) *Exhortation*	*Timing*	*Kapcha*
Control	77.4	0.734	0.143	0.002**
Exhortation	76.4		0.057	< 0.001***
Timing control	83.9			0.076
Kapcha	90.2			

[21]Assuming that the true difference in the proportion of attrition between the *Control* and *Exhortation* treatment was 3.4%, we would need 620 observations in both treatments to have an 80% chance of detecting it.

[22]In many survey situations such as surveying current members of an organization, maximizing the response rate is a good strategy. However, in academic research, especially behavioral economics and psychology, weeding out non-serious respondents may be desirable.

[23][16] discusses how excluding data based on whether people fail the instructional manipulation check may lead to a non-representative population. This is a concern that should be considered by the researcher, but is unlikely to be relevant unless the non-representative subset of workers would bias the study.

Table 4 illustrates that the *Control* and *Exhortation* treatments differ significantly from the *Kapcha* treatment.[24] The difference between the *Timing Control* and *Kapcha* was almost significant ($p = 0.076$). We suspect this difference is real but we most likely do not have enough data to detect it.[25]

Table 5: IMC pass rate (in %) explained by treatment and other covariates

($N = 727$)	without controls b (se)	with controls b (se)
Treatment		
Exhortation	-1.0 (4.4)	-1.0 (4.4)
Timing Control	6.6 (4.2)	7.3 (4.2)
Kapcha	12.8** (4.0)	13.0*** (3.9)
Gender (male)		-7.7* (3.1)
Age (26-35)		11.4** (4.0)
Age (36-45)		16.3*** (4.3)
Age (over 45)		17.1*** (4.6)
Completed college		4.2 (2.9)
Reported motivation		1.9 (1.1)
# words in feedback		0.3*** (0.1)
Need for cognition		3.8 (2.5)
Break for $\geq$ 2min		-13.1 (7.7)
Other covariates		✓
Intercept	77.4*** (3.2)	27.1 (14.3)
R^2	0.020	0.165

*p < 0.05, **p < 0.01, ***p < 0.001

Table 5 demonstrates that the only experimental treatment which significantly impacts the IMC pass rate is the *Kapcha* fade-in treatment, increasing the pass rate by 12.8± 4.0% ($p < 0.01$). Controlling for demographics[26] makes the effect more statistically significant while leaving the estimate unchanged ($p < 0.001$). As expected, controlling for demographics also significantly improves the overall fit of the model, as measured by R^2. Demographic factors that affect the IMC pass rate are discussed in section 3.4.

Are the higher IMC pass rates in those groups simply the result of "less serious" respondents removing themselves from our sample? We investigate whether the higher attrition rates in the *Kapcha* treatment can explain the differences in IMC pass rates we have been exploring.

Under the most conservative assumptions, we assume that all of the additional workers who left the *Kapcha* would have stayed and subsequently failed the IMC. More specifically, we assume that the differential attrition between the *Kapcha* and the *Control* treatments (13.3% - 5.6% = 7.7%) stayed and fail the IMC (188×7.7% = 14 new failing workers). This lower IMC pass rate would become 83.05% which is greater

[24]two-sample, two-tailed z-tests

[25]If the true proportions were equal to the means that we observed, we would need 492 observations in each treatment to have an 80% chance of detecting it.

[26]The other covariates in table 5 represent covariates that were not significant and jointly, barely significant. These include frequency of survey-taking on MTurk, hours per week on MTurk, earning per week on MTurk, task day of week, task hour of day, reported multitasking behavior, minutes since HIT was listed, and average time spent on first 30 questions.

than the 77.4% pass rate of the *Control* ($p = 0.09$, one-tailed two-sample z-test). Under these conservative assumptions, attrition can only explain 5.65% of the total 12.8% effect, or 44% of *Kapcha*'s success. It is reasonable to assume that the *Kapcha* adds an additional boost beyond merely annoying people until they leave.

3.3 Finding Larger Effects in the Economic Behavior Questions

Assuming that the subtle word changes from questions A and B cause real differences, we hypothesize that the largest effects in both questions will be found within the *Kapcha* treatment which is designed to force people to pay close attention to the words in the question text.

Tables 6 and 7 show the effect of subtle word changes on the price subjects pay for a soda when it comes from a "fancy resort" and the increased intent to go to a game whose tickets were "paid handsomely" for as opposed to received for free.

Overall, workers would pay \$0.24 more for sodas bought at a "fancy resort" over a "run-down grocery store" ($p < 0.01$). As expected, in our *Timing control* and *Kapcha* treatments, this effect is stronger than average and is largest and highly statistically significant in the *Kapcha* treatment. Controlling for demographics leaves this result unchanged, but substantially improves the overall fit of our model.

Overall, workers who paid for the football ticket were more likely to go than workers who received the ticket for free with a difference of 0.37 intention units on a nine-point scale ($p < 0.05$). The effect was not robust when controlling for demographic variables.

The effect of questions B's phrasal change is estimated to be between 0.29 and 0.37 depending on whether we control for demographics. In no individual treatment are there significant differences both with and without control variables. That said, the *Timing control* and the *Control* treatment appear to have a larger effect. However, these results are barely significant and vacillate upon the introduction or removal of the demographic controls. We consider these effects to be spurious and conclude that the standard errors in each treatment are too large (approximately 0.31 to 0.44 depending on the treatment) relative to the hypothesized effect size. Most likely, there was not enough data to detect the small effects given the considerable spread in responses.

The analysis of question A lends legitimacy to the *Kapcha* treatment's power to reduce satisficing. However, we were unable to draw conclusions from the responses to question B. We again note that both investigations were underpowered. We hope to get more data in the future so we can use the effects found in questions A and B to proxy for satisficing behavior.

3.4 Observations on the MTurk Survey-taking Population

MTurkers Beat Stanford and NYU Students

We compare the IMC pass rates from [16] with our data. In [16], using $n = 213$ New York University undergraduates, the IMC pass rate was 54%, which is lower than our *Control* group ($n = 167$) with a pass rate of 77.3% ($p < 0.001$). This *Control* treatment pass rate is similar to the 82.5% pass rate in [16] during administration of a paper and pencil exam using $n = 336$ Stanford university undergraduates who were believed to be "motivated" (because they were interested in either a major or minor in psychology).

Demographic and Behavioral Drivers

Which demographic groups paid the closest attention to our survey? In table 5, we find that women on average pass the IMC $7.7 \pm 3.1\%$ more often than men ($p < 0.05$). We find that older workers do better than younger workers; 26–35 year olds pass $11.4 \pm 4.0\%$ more often ($p < 0.01$); 36–45 year olds pass $16.3 \pm 4.3\%$ more often ($p < 0.001$); and workers over 45 years of age pass $17.1 \pm 4.6\%$ more often ($p < 0.001$). We also find that workers who completed college pass 4.2% more often (this result was nearly significant).

Two of our variables which measure respondent engagement, the NFC and self-reported motivation, are not significant when included together and with the number of words in feedback. However, these variables are both highly significant (when included only with the indicator variables for treatments).

Surprisingly, the worker's average number of hours worked on MTurk per week, the average earnings on MTurk per week, the reported level of multitasking, nor the frequency of survey-related tasks were significant in predicting IMC pass rate.

Feedback

The final significant relationship found was the number of words written as feedback (question #31). We find that for each word of additional feedback, the probability of passing increased by $0.3 \pm 0.1\%$. A one standard deviation change in the number of words of feedback (equal to 28.2 words) is associated with an 8.5% increase in the IMC pass rate even controlling for other measures of engagement. Therefore, the length of a free response can be used as a proxy for survey engagement, a result also reported by [1].

We also did a small experiment studying how to incentivize feedback. We varied how much we offered respondents for providing feedback (offering either one, zero or five cents). The average feedback in the group without a bonus is 28.3 words. Compared with an unpaid bonus, paying a one cent reward garners 5.0 more words on average ($p < 0.05$) and the five cent bonus garners 7.6 more words ($p < 0.01$). Further, we could not reject the hypothesis that the two effects were equal ($p = 0.364$). This indicates that paying a minimum bonus of one cent elicits almost as lengthy feedback as paying almost five times that much (and roughly half the value of the full HIT). Interestingly, although we expected the *Exhortation* group to provide more feedback since they were reminded that they were participating in the study, neither that group nor any other treatments received significantly more feedback (although the groups were jointly significant at the $p < .05$ level).

Furthermore, the workers are eager to give feedback. Researchers can rapidly pilot their studies and get real-time feedback on how they are perceived by survey-takers.

Asking feedback also gave us a wealth of insight into how survey respondents perceived our various treatments including the *Kapcha*. One danger of setting the reading speed too slow for fast readers was illustrated by this worker from Colorado Springs, CO:[27]

[27]We recorded each worker's IP address which allowed us to determine their location.

Table 6: Question A: Increase in willingness to pay for a soda due to subtle word changes involving whether source of soda was a fancy resort or run-down grocery store (with and without other controls)

	Overall	*Control*	*Exhortation*	*Timing control*	*Kapcha*
No controls					
"fancy" b (se)	0.239**	0.080	0.072	0.303	0.543***
	(0.086)	(0.173)	(0.176)	(0.176)	(0.148)
R^2	0.011	0.001	0.001	0.015	0.077
With controls[a]					
"fancy" b (se)	0.249**	0.143	-0.021	0.319	0.533**
	(0.089)	(0.180)	(0.184)	(0.180)	(0.182)
R^2	0.083	0.210	0.185	0.254	0.205
N[b]	714	163	201	190	160

*p < 0.05, **p < 0.01, ***p < 0.001
[a] Includes same controls as table 5
[b] We excluded 13 prices that were not numbers between $0 and $10

> "Text needs to be instantaneous. No apparent reason for it to appear slowly other than to aggravate the participant."

However, this comment from another worker in Detroit, MI illustrates the intended purpose:

> "I didn't enjoy the way the words scrolled slowly, as I read fast, but in its defense the slow scrolling words lead me to pay closer attention to what I was reading and skim less."

In addition, many workers are survey-savvy and eager to offer design suggestions such as this worker from San Bernardino, CA who is also familiar with Likert scales:

> "I liked the situational question about the soda cost... I do not like the black background color, it hurts my eyes when contrasted with the white. Would prefer a 7 point likert scale if possible."

4. DISCUSSION

The main goal of our study is to investigate a survey platform that reduces satisficing across the board. We propose the idea of *Kapcha*, a method which involves slowing people down by fading-in the question text, thereby accentuating each word. We have found evidence that *Kapcha* has the potential to reduce satisficing in online surveys. We then open-source the platform (see Appendix A) so that the surveyor can simply "plugin" the platform and be confident of obtaining more accurate results.

MTurk workers that participated in a survey task employing the *Kapcha* passed an *instructional manipulation check* about 13% more often than those who were given a standard survey and it is reasonable to assume that this pass rate can be used as a proxy for general satisficing behavior. At most, only 44% of this effect can be explained by a higher proportion of people leaving the *Kapcha* survey task.

The treatment where we merely exhorted the participant to pay more attention had no significant effect on satisficing. The treatment that imposed a waiting period but did not accentuate the words, did better than the standard survey group, but the difference was not significant.

Upon analyzing demographic data, we find the segment of workers least likely to satisfice are females over the age of 26 who leave thoughtful feedback.

We must also emphasize that the trick question was very difficult and requires carefully reading the fine print.[28] As a testament to the quality of work on MTurk, we find it absolutely incredible that even in the *Control* treatment, people pass the trick question with such high proportion.

5. FUTURE DIRECTIONS

Our study indicates that using *Kapcha* can significantly increase the amount of attention respondents give to reading directions and answering questions. For future research, we would like to study how the *Kapcha* is affected by other variables such as levels of motivation or monetary incentives as well as among different populations. We would also like to conduct further experiment with how the *Kapcha* could be optimized using principles of psychology and perception so as to draw the attention of respondents. Finally, we would like to design a survey task that is deliberately designed to impose a high cognitive burden and cause respondents to satisfice. Testing the *Kapcha* under these circumstances will provide a clearer picture of its power.

Kapcha may be moderated by other variables

For one reason or another, the *Kapcha* may be more effective on certain populations. For instance, the degree to which a respondent pays closer attention when being forced to wait, or when text is faded in, may differ by language or culture. A study drawing participants from various countries may elucidate its differential effectiveness.

Apart from interactions with demographic variables, the effectiveness of the *Kapcha* may vary with monetary or non-monetary incentives.

For example, can you pay people to pay more attention? If people are paid higher monetary awards, does that reduce the advantages of using a *Kapcha*? It could be that incentives simply cannot induce people to pay more attention beyond a certain point and that the only way to increase attention to the highest levels is through attention-grabbing techniques.

[28] One of the authors gave it to colleagues in their department and each of them failed.

Table 7: Question B: Increase in intention to attend the football game due to subtle word changes involving whether the participant paid for or received a ticket for free (with and without other controls)

	Overall	*Control*	*Exhortation*	*Timing control*	*Kapcha*
No controls					
"paid"	0.370*	0.716	0.066	0.797*	0.039
b (se)	(0.176)	(0.365)	(0.314)	(0.376)	(0.381)
R^2	0.006	0.024	0.000	0.024	0.000
With controls[a]					
"paid"	0.285	0.981*	0.150	0.726	-0.026
b (se)	(0.177)	(0.443)	(0.330)	(0.385)	(0.378)
R^2	0.066	0.203	0.162	0.256	0.322
N	727	168	203	193	163

*p ¡ 0.05, **p ¡ 0.01, ***p ¡ 0.001
[a] Includes same controls as table 5

Though not statistically significant, it appears that telling respondents that their answers will be used for research (our *Exhortation* treatment) motivates people to complete our survey at higher rates. This is most likely because reminding them about the survey's research value imbues the survey with a sense of meaning (a similar result was found in [4]). Therefore, it may be wise to insert an exhortative statement into the *Kapcha* to get an "added boost."

The Kapcha appearance should be optimized

Further, we would like to experiment with the particulars of the *Kapcha* presentation. In our experiment, we used white text on a black background and faded the words in at 250 words per minute. Does the choice of color schemes and text font matter?[29] What is the speed at which words should fade-in to optimize attention to our survey?

Research suggests that forcing a person to direct their gaze toward an area correlates highly with the attention they pay to that area [8] and the psychology of perception is ripe with many other examples of how color, contrast, and movement could be used to draw attention.

We must admit that many of our faster reading respondents in the *Kapcha* group expressed that they did not like the unfamiliar method of fading-in survey questions (34 of 163 respondents left negative feedback compared with 7 positive feedbacks). Although it had its desired effect of increasing the attention people paid to the survey questions, we certainly would like to further calibrate the *Kapcha* so as to slow respondents down without annoying them.

We propose a survey task that measures satisficing more generally

The IMC and the behavioral questions are both noisy and incomplete measures of satisficing. Using only these response variables as proxies for satisficing is a weakness in our present study.

We propose to create a survey task that has several measures of satisficing in order to demonstrate that the *Kapcha* can reduce satisficing in the broadest context.

We will review the findings from the literature of survey response psychology (e.g. [15], [13], [22]) which provide guidance for how to design surveys to minimize participant satisficing. Using these principles, we will reverse-engineer a survey that is *deliberately* constructed so that respondents are *likely* to satisfice. We will then see how well the *Kapcha* prevents satisficing even under the most difficult of circumstances.

To offer an example, [14] provides a framework for how respondents satisfice depending on the *structure of questions*, the survey's *difficulty*, the *respondent's ability*, and the *respondent's motivation*. Three commonly cited examples of satisficing due to question structure are *response order effects* whereby people choose the first answer of surveys, *no opinion filters* whereby people who are lazy will sooner choose "no opinion" than take the time to think of what their opinion is, and *acquiescence bias*. where respondents are more likely to choose "agree" if the choices are "agree or disagree". The difficulty of a survey can be related to the "readability" of the survey questions (higher readability implies shorter question length and basic vocabulary [3]). The respondent's motivation may be related to how meaningful they perceive the task to be. Presumably, higher ability respondents and respondents who are motivated would also tend to satisfice less.

If the *Kapcha* is found to be effective here, we will be confident that the *Kapcha* method prevents satisficing under very general conditions.

Data Sharing

We cross-validated some of the self-reported demographic information using data provided by Panos Ipeirotis from [11]. 23 people who reported their age in our survey also reported their date of birth in [11]'s survey. In all but one case, the age and date of birth were consistent. This offers evidence that even over a time period of more than 6 months, time invariant demographic data can be reliably collected on separate occasions. Broadly speaking, MTurk workers seem to be honest in sharing their personal information.

Data sharing among academics using MTurk provides not only the possibility of validating data, but also of using demographics or other covariates from one study as controls in others. For example, in our present study, we evaluated respondent's Need for Cognition which could be a useful control variable in other studies. In many cases, it may be highly useful to match demographic and other behavioral characteristics as a way to increase precision without us-

[29]Many respondents complained that the white-on-black background was distracting which may have affected how well the Kapcha worked.

ing the limited time of respondents. Using data from other studies is especially beneficial in the case of natural field experiments where the researcher will not want insinuate that the task is an experiment.

We propose that researchers agree on a central, shared repository of data related to the MTurk workers and offer an API for easy access.

6. ACKNOWLEDGMENTS

The authors wish to thank Larry Brown, Persi Diaconis, David Gross, Susan Holmes, Daan Struyven, Nils Wernerfelt, and Adi Wyner for helpful discussion and comments. Both authors also acknowledge support from the National Science Foundation in the form of Graduate Research Fellowships. We would like to thank Aptana Inc. for providing a hosting platform for the experimental software.

7. REFERENCES

[1] A. Bush and A. Parasuraman. Assessing response quality. A self-disclosure approach to assessing response quality in mall intercept and telephone interviews. *Psychology and Marketing*, 1(3-4):57–71, 1984.

[2] J. Cacioppo, R. Petty, and C. Kao. The efficient assessment of need for cognition. *Journal of Personality Assessment*, 48(3):306–307, 1984.

[3] J. Chall and E. Dale. *Readability revisited: The new Dale-Chall readability formula*. Brookline Books, 1995.

[4] D. Chandler and A. Kapelner. Breaking monotony with meaning: Motivation in crowdsourcing markets. *University of Chicago mimeo*, 2010.

[5] L. B. Chilton, J. J. Horton, R. C. Miller, and S. Azenkot. Task Search in a Human Computation Market. 2010.

[6] M. P. Couper, R. Tourangeau, and T. Marvin. Taking the Audio Out of Audio-CASI. *Public Opinion Quarterly*, 73(2):281–303, May 2009.

[7] T. DeMaio. Social desirability and survey measurement: A review. *Surveying Subjective Phenomena*, 2:257–282, 1984.

[8] J. Hoffman and B. Subramaniam. The role of visual attention in saccadic eye movements. *Perception and Psychophysics*, 1995.

[9] S. Holmes and A. Kapelner. Quality assessment of feature counting using a distributed workforce: Crowd counting a crowd (in progress). 2010.

[10] J. J. Horton, D. G. Rand, and R. J. Zeckhauser. The Online Laboratory: Conducting Experiments in a Real Labor Market. *SSRN eLibrary*, 2010.

[11] P. Ipeirotis. Demographics of Mechanical Turk. CeDER working paper CeDER-10-01, New York University, Stern School of Business., Mar. 2010.

[12] J. A. Krosnick. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5(3):213–236, May 1991.

[13] J. A. Krosnick. Survey research. *Annual Review of Psychology*, 50:537–67, Jan. 1999.

[14] J. A. Krosnick. The threat of satisficing in surveys: the shortcuts respondents take in answering questions. *Survey Methods Centre Newsletter*, 2000.

[15] J. A. Krosnick, S. Narayan, and W. R. Smith. Satisficing in surveys: Initial evidence. *New Directions for Evaluation*, 1996(70):29–44, Mar. 1996.

[16] D. M. Oppenheimer, T. Meyvis, and N. Davidenko. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4):867–872, July 2009.

[17] G. Paolacci, J. Chandler, and P. G. Ipeirotis. Running Experiments on Amazon Mechanical Turk. 2010.

[18] H. Simon. A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, 1955.

[19] A. Sorokin and D. Forsyth. Utility Data Annotation with Amazon Mechanical Turk. *First IEEE Workshop on Internet Vision, CVPR 08*, 2008.

[20] S. Taylor. Eye movements in reading: Facts and fallacies. *American Educational Research Journal*, 2(4):187, 1965.

[21] R. Thaler. Mental Accounting and Consumer Choice. *Marketing Science*, 4(3):199 – 214, 1985.

[22] R. Tourangeau, L. Rips, and K. Rasinski. *The Psychology of Survey Response*. Cambridge University Press, 2000.

[23] L. Von Ahn, M. Blum, N. Hopper, and J. Langford. CAPTCHA: Using hard AI problems for security. *Advances in Cryptology EUROCRYPT 2003*, 2003.

APPENDIX

A. TURKSURVEYOR: AN OPEN-SOURCE EXPERIMENTAL PLATFORM

We would like to introduce "TurkSurveyor", an open-source experimental system designed for running surveys (or survey-based experiments) on Amazon's Mechanical Turk. TurkSurveyor is written in a mixture of Ruby (on Rails), HTML, CSS, and Javascript and is available under the MIT license at http://code.google.com/p/turksurveyor/ and includes an instruction manual. The goal of its development is to have a simple push-button system which allows one, with a minimum of customization, to use MTurk to collect data for a custom survey.

B. REPLICATION

At http://danachandler.com/kapchastudy.html, you can find the source code, the raw data, and the analysis used to run this study.