4

METHOD

In this chapter, I outline a general research program and associated substantive and experimental hypotheses. Following, I describe the use of Amazon Mechanical Turk for the collection of behavioral data, general procedure followed, and summary of the population sample observed.

4.1 RESEARCH PROGRAM

There is little question that interaction with computers change what we know and believe. Usability has long been concerned with the fit between user "mental model" and the design of user interfaces. Typically, such studies include the use of qualitative research techniques such as cognitive walk-throughs, eye-tracking, and card sorting. Such studies are often task-oriented and situated to address the use of specific products.

Dan Saffer, in his book *Microinteractions* (2013), touches upon user interaction with user interface components, or features, that perform only one small task. For instance, a login screen embodies a self-contained interaction that may be quite small, but extremely important for user experience. This microinteraction has a clear start and end. As Saffer notes, some triggers to enter a microinteraction are user-initiated, and some may be system-initiated. Regardless, I believe there is a clear parallel to conversational discourse discussed in the previous chapter: users exchange information interactively.

This dissertation examines several microinteraction in the domain of online behavioral advertising. Of particular interest are phenomena relating to user confusion in the context of online interaction.

The substantive hypothesis for questions asked in this dissertation is that *some user confusion stemming from online behavioral advertising is caused by error in discourse understanding*. I contend that by regarding user interaction with graphical user interfaces as a form of discourse, we can account for some sorts of user confusion that may be difficult to see or may otherwise go un-noticed.

The next section summarizes specific research hypotheses for this dissertation. The research approach used to examine these hypotheses is *randomized experimentatal*. The purpose for adopting such an approach is to determine causality for some sorts of user confusion. Each research hypothesis is paired to some aspect of discourse theory. Therefore, Chapters 5, 6, and 7 more fully motivate the connection between theory and practice prior to a presentation of results.

4.2 DESIGN OF EXPERIMENTS

The substantive hypothesis (user confusion stems from errors in discourse understanding) implies a set of research hypotheses, which in turn imply individual study hypotheses.

RESEARCH HYPOTHESIS 1:
: Graphical user interfaces have properties comparable to other forms of linguistic communication.

RESEARCH HYPOTHESIS 2:
: Graphical user interfaces can cause users to make faulty pragmatic inferences.

RESEARCH HYPOTHESIS 3:
: A pragmatic inference derived while interacting with a graphical user interface may be affected when given immediate and direct feedback.

Each of the three sets of experiments described in Chapters 5-7 support one more more of the research hypotheses above. Specific research manipulations are described below. The number in the left column refers to the experiment number.

Research manipulation (by experimental condition):

1A: Suppose that pragmatic inference only occurs in linguistic (text or speech) interaction. A graphical user interface is used in the experimental condition whereas text is used in the control condition.

1B: Suppose that system feedback adds information to a situation model. Feedback is used in the experimental condition while none in the control condition.

2A: Suppose images and icons are objects that are both indexical. Images and icons are combined such that contrastive relations exist in the experimental condition while none exists in the control condition.

2B: Suppose advertising images are implicitly indexical. Explicit indexical element appears in the experimental condition while none exists in the control condition.

3: Suppose that a user designs utterances taking into account potential listeners. Listener roles are explicit in the experimental condition while implied in the control condition.

Finally, the table below summarizes experimental designs. Each should be considered a pilot study. That is, experiments described are new and there are no other results against which to compare. The goal is to apply difference statistics to determine whether the variables described are independent (null hypothesis) or dependent. A relation between variables indicates some degree of predictiveness and causality. We can use information about causality to guide, or motivate, a strategy for reducing the possibility of miscommunication.

Table 1: Experiment Design

1A	2x2x2 between group posttest-only design where the control group is presented with a set of textual expressions and asked to answer questions about their meaning. Treatment groups are presented with either textual expressions or a dialog box expressing the same set of choices and asked the same questions. Independent Variables: Modality, Deontic force, Attitude toward privacy. Dependent Variable: Pragmatic implicature
1B	Between group posttest-only design where two groups are presented with a cookie banner and later asked about whether or not they believe the website placed "cookies" in their browser. The treatment group is presented with feedback about the consequence of their action / non-action following presentation of the banner. Independent Variable: Feedback. Dependent Variable: Pragmatic implicature.

2A	Between group posttest-only design where five groups are presented an advertisement in the context of a webpage and asked to identify hyperlinks. Treatment groups are presented an advertisement with embedded image icons at four levels (known icon + different company, unknown icon, known call-to-action (CTA), "DAA opt-out icon") while the control group is presented with an embedded image from the same company as the advertiser. Independent Variable: Icon type. Dependent Variable: Indexicality of icon.
2B	Between groups posttest-only design where three groups are presented an advertisement in the context of a webpage and asked to identify hyperlinks. Treatment groups are presented an advertisement at two levels (iconic CTA, textual CTA) while the control group is presented with an advertisement with no CTA. Independent Variable: Modality of CTA. Dependent Variable: Click target.
3	Between group repeated measures posttest-only design where a control group is asked to respond to a survey containing questions relating to activities (using questions drawn from Acquisti, John, and Loewenstein (2012); embarrassing, socially / ethically questionable, illegal). Participants are notified in a privacy statement that there may be ad trackers on the site. The treatment group receives exactly the same notification and survey. However, tracker presence is indicated in a visual display throughout the participant's session. Independent Variable: Visual presence. Dependent Variable: Propensity to respond in the affirmative to engaging in specific behaviors.

Of course, not every usability problem needs be tested using experimental methods. But there are two good reasons to do so here:

1. No causality has yet been determined for presumed cases of mis-communication studied here. In fact, it may not even be apparent that there is a problem at all.

2. Though linguistic theory offers explanation for the sort of miscommunication described in experiments here, there are no earlier studies with which to compare. This dissertation intends to make such a relation clear and provide a path for future exploration.

Each of the research hypotheses above has bearing on current practice in online behavioral advertising as represented in technical specifications of existing or proposed standards. Though self-regulatory bodies for OBA suggest that specific principles are adhered, effectiveness is in question. Marketing advocates affirm that self-regulatory practices are effective (Tribal Fusion, 2012). However, research studies by policy advocates (e.g., Komanduri, Shay, Norcie, & Ur, 2012; Mayer & Mitchell, 2012) have raised issues of industry non-compliance. Even if compliance were universal, usability studies have uncovered inadequacies in terms of communicative efficacy (Hastak & Culnan, 2010; Leon et al., 2012; Ur, Leon, Cranor, Shay, & Wang, 2012). The benefit of the approach in this dissertation is the ability to test technical specifications, and ascertain effectiveness in context.

One of the particular challenges of the randomized experimental approach is the need for a larger number of participants than one might expect from a traditional usability experiment. Increasingly, psychologists and linguists have come to rely on the large pool of volunteers available via the Amazon Mechanical Turk (AMT) crowdsourcing platform. The next section addresses both advantages and disadvantages of AMT for experimentation. Characteristics of the user sample studied have particular bearing on the research validity of experiments presented in this thesis.

4.3 MECHANICAL TURK AS A PLATFORM FOR HUMAN INTELLIGENCE TASKS

The original Mechanical Turk constructed in the 18th century was a hoax. In all appearance, he was a mechanical man dressed in an turban and robe, seated behind a large cabinet. Despite being mechanical, he played chess like no other machine could. Of course, unknown to most at the time, inside the cabinet was a tiny space where a real human played the game by controlling the automaton's arms on the chessboard.

By contrast, Amazon's Mechanical Turk (AMT) is no hoax. From a simple command line interface, requester's can create

"human intelligence tasks" and farm them out to a sea of workers. Within seconds, a task may be accepted, completed, and remunerated without the requester ever knowing or communicating with the invisible worker. AMT shares a simple design concept with the original Turk: a human is in the machine. The idea is a "job posting board" where human intelligence is needed. Any of a variety of "Human Intelligence Tasks" (HITs) can be posted along with small monetary rewards ranging from cents to dollars. AMT is a prototypical example of crowdsourcing — outsourcing to a large, undefined community of people (i.e., crowd).

4.3.1 *The AMT Marketplace*

In 2006, Wired magazine published the influential *The Rise of Crowdsourcing* (Howe, 2006), which tells the story of the rise of the amateur on the web. Highlighted, is the example of a stock photo site iStockphoto which created a marketplace for amateur photographers. Professional grade camera technology at affordable prices combined with the ability to upload content to the Internet, search and categorization, and micro-payments changed the market for photographers and photo purchasers. Over the last decade, the power of many users over distributed networks has changed the world. Today, FaceBook and eBay could not exist without the contributions of their users.

The irony is, of course, that by engaging in the creation of user-generated content on the web, people have become part of the machine. This has led to the well-known technology humanist Jaron Lanier to rail against the de-humanization of the web. His position is — to design applications that treat humans as machines devalues both people and content "making ourselves into idiots" (Lanier, 2006). Lanier particularly dislikes Wikipedia which creates a false sense of authority behind information by removing any connection between the real author of the information as well as a subjective context for the interpretation of content (Lanier, 2010).

Nevertheless, crowdsourcing has become a viable economic model on the web. More than a million workers login to crowdsourcing platforms to complete short tasks for pay-per-task compensation (Munro & Tily, 2011). And who is the invisible worker? As reported on The Mechanical Turk Blog in August of 2012, "turkers" comprise a workforce approxi-

mately 500,000 persons (at any one time) across 190 countries. In an analysis of the AMT marketplace over multiple studies, Mason and Suri (2011) report that the majority of turkers reside in the United States or India and tend more likely to be female (55%) with a median average age of 30 and, on average, earning 30K per year. According to Ipeirotis (2010a; 2010b), workers in the US are more likely to use AMT as a secondary source of income while those in India use it as a primary source of income. Workers are self-selected on the marketplace, working for cash payments in US dollars and Indian rupees. All studies agree that workers are represented across a wide range of ages, ethnicities, countries, languages and income.

The AMT marketplace has been active since 2005. Among the most well-known documented instances of use, turkers were asked to help search through sections of satellite imagery to flag images with anomalies. The hope was that turkers might be able to help pinpoint the 2007 crash site of aviator Steve Fossett. This event so permeated the media at the time, I signed up to become a turker in order to participate. In the end, Fossett was not found in the region examined by turkers. In fact, there was some criticism toward the use of turkers for annotating satellite imagery. With little understanding of what constitutes anomaly in satellite imagery, the usefulness of generated leads were considered by some to be questionable.

Since then, AMT has been used for a variety of tasks, some of which call for specific skill or expertise, such as knowledge of a foreign language. More common are tasks that involve transcription, rating, editing, and image annotation. The AMT website provides a basic web interface for loading HITs, though it is also possible for requesters to direct workers to an external website or directly embed that website as an iframe on the AMT site.

Workers themselves comprise a community that interact on public forums, such as *Turker Nation*, and through networks such as Twitter. Once on the AMT website, workers track their earnings (managed by Amazon), status (HIT success), qualifications, and the ever revolving database of available tasks. Figure 1 is a screengrab taken from the AMT developer sandbox, a place for requesters to test tasks before deploying on the AMT website.

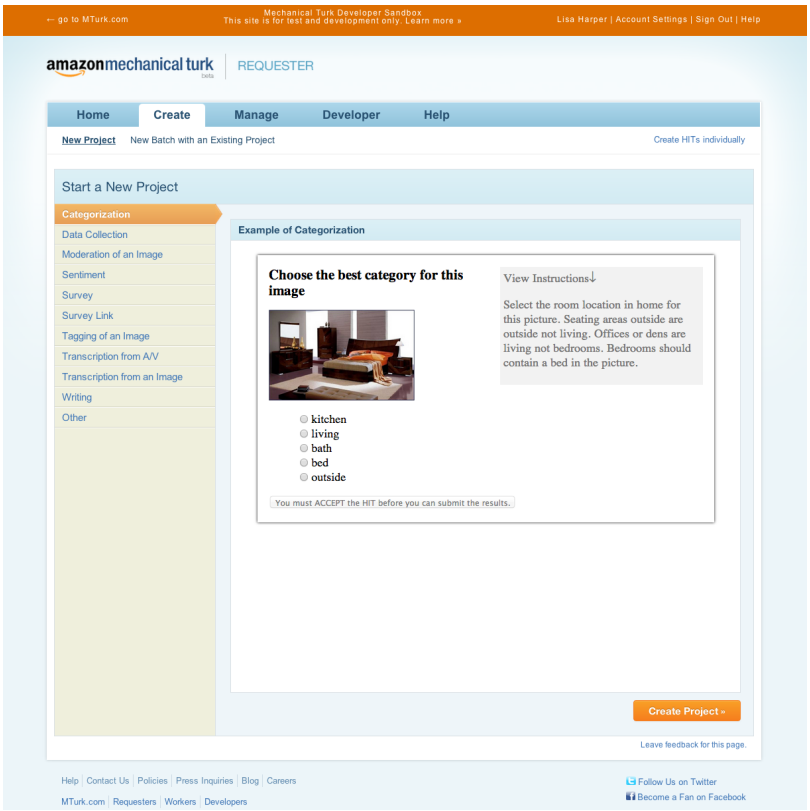

Figure 1: AMT Requester Website

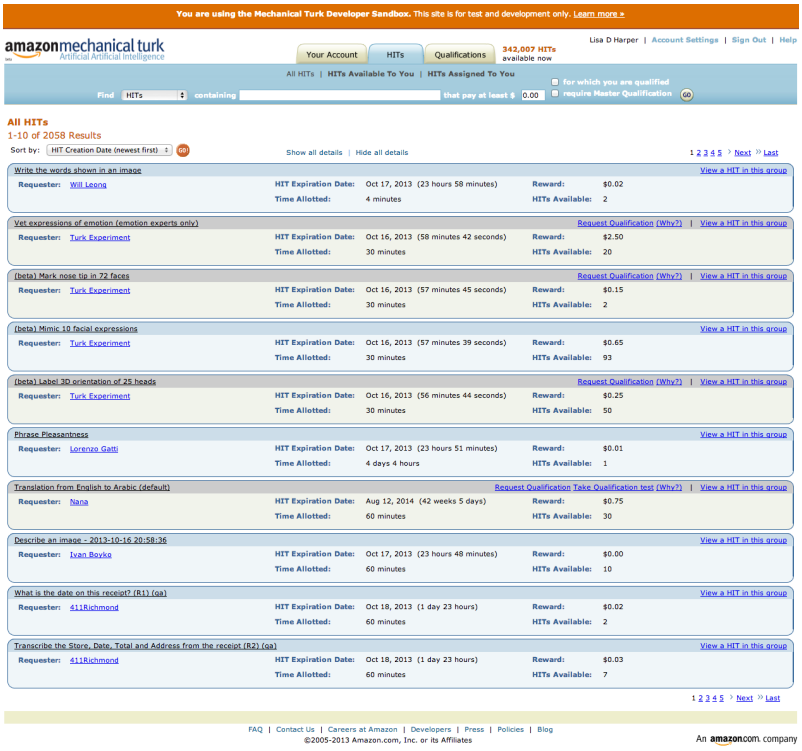

Figure 2: AMT Worker HITs

Requesters define the task, number of tasks available, payment, qualification, and expiration date. Requesters must provide a validated US bank account and address in order to initiate tasks. Some basic worker qualification types are provided by Amazon out-of-the-box: worker HIT success and demographic region are commonly used by requesters. However, requesters also use qualifications as a mechanism for training workers. By requiring workers to "train" and go through testing, workers with special skills may be utilized in future tasks.

Because the worker success metric is often used as a filter for ensuring higher quality work — and as a means for obtaining higher paid HITS — workers take care to protect their rating. This means that requesters are also held accountable in their dealings with workers, via informal ratings over social media channels. Though some HITs may be paid automatically, it is fairly common for requesters to do quality control before accepting HITs. HIT rejection has negative consequences and is not taken lightly by the turkers.

There are a number of obvious advantages to using AMT. The workforce is roughly half a million strong and available around the clock. They represent a wide demographic range, many quite well educated. Amazon manages and automates payments and ensures the pool of workers abides by its terms of use. Mason and Suri (2011) observed that, for experimental research, a key advantage is faster iteration between hypothesis formulation and testing. However, there exists a number of real concerns and issues with crowdsourced experimentation. These are discussed below.

4.3.2 *General Concerns with the Use of AMT for Experimental Research*

First, and foremost, AMT is an Internet-based platform. Reips (2006) noted advantages as well as disadvantages to experimentation over the Internet (relative to the first four bullet points below). However, there are also number of other issues particular to Mechanical Turk.

1. **User Fraud.** There is the possibility of worker "gaming" through multiple submissions. The Turk Requester's Blog

[1] gives an example from the Romanian Mturk Forum, in which on every page (on a total of more than 300), are discussions of how to break Amazon's terms of service. As a result, Amazon adjusted their policy to an invitation-only registration for International workers (Chiarella, 2013).

2. **Self-Selection.** Users self-select by reviewing titles, payment, time expected, and instructions. Unless controlled, users are generally able to preview eternal websites (or embedded iframes) to decide whether they wish to accept a given HIT.

3. **Drop-outs.** AMT workers have the ability to return or abandon HITs without affecting their reputation. However, requester's don't automatically see which workers accept — though don't complete — assignments, so its difficult to gauge dropouts.

4. **Mis-communication.** Reduced communication between requesters and workers may mean that directions are not well understood or followed. This can be difficult to detect and debug.

5. **Screening.** Amazon provides no general mechanism for screening aside from a general mechanism for filtering via qualifications. Of note, is the need to screen for demographic criteria and prior participation. This study has the particular need for the latter: excluding workers who have participated in one experiment from participating in another. The mechanism that AMT provides is exclusion by assignment: once a worker has accepted an assignment, that particular assignment is no longer available. For example, Experiment 2A of this dissertation required at least 300 subjects. Once a worker accepted this HIT the other 299 assignments were no longer available. But, without cross-experimental controls, it would be possible for some worker to accept a HIT from a closely related experiment, potentially biasing results.

6. **Compensation.** The very fact that AMT is fee-based places obligation on both the requester and worker. Mason and Watts (2009) investigated the relationship be-

1 http://turkrequesters.blogspot.com/2013/01/the-reasons-why-amazon-mechanical-turk.html

tween compensation and performance in two experiments on AMT. They observed no interaction between difficulty of task and compensation on performance. They also found that increasing compensation alone did not improve accuracy. But how workers were paid (pay-per-word, pay-per-puzzle) did have an impact on output and accuracy.

While it is common for workers to accept tasks for pennies each, some workers find by working many low paying tasks in quick succession, they are able to earn close to minimum wage.[2] More typically they earn much less (Paolacci, Chandler, & Ipeirotis, 2010). Though most workers appear attracted by the earnings potential, others participate for reasons such as boredom, fun, curiosity, and even education (Behrend, Sharek, Meade, & Wiebe, 2011).

The studies in this dissertation limit participation to workers with a high HIT acceptance rate. From discussion on turker forums and blogs (e.g., *Turker Nation*, *mTurk Forum*, and *Turkkit-Reddit*), its clear that these workers expect to be paid more than other workers. And many take this work quite seriously.

7. **Quality and Reliability.** Beyond recruiting from the most reliable workers available, how much should quality be of concern? From prior crowdsourcing experiments on Amazon Mechanical Turk (AMT), quality is a valid concern (Callison-Burch & Dredze, 2010; Gormley, Gerber, Harper, & Dredze, 2010). In a study of translation tasks, researchers consistently receive poor and noisy translations including blank annotations, misspellings, copy-pasting of machine translations, and downright cheating (Ambati & Vogel, 2010). Having workflows for the filtering of noisy data has proven vital in this environment. However, it is feasible to filter noisy judgements of non-experts depending on the task: in an image annotation task Nowak and Rüger (2010) found that, while agreement between experts and non-experts varies depending on the measure used, its influence on image ranking as a whole, is minimal. Other sorts of studies echo this finding including labeling text with emotion

2 See discussion at http://mturkforum.com/showthread.php?2744-How-much-do-you-earn-per-hour, for example.

(Snow, O'Connor, Jurafsky, & Ng, 2008), search relevance judgements (Alonso & Mizzaro, 2009), and more.

8. **Presentation Consistency.** All Internet experiments conducted in the wild suffer the need for exceptional attention to cross-browser effects and robust services. If the experiment relies on everyone being able to see stimuli at the same resolution and in the same manner, then both scripts and screens need to be tested carefully, both across platforms and browsers. Furthermore, hosted services require reliable access by hundreds simultaneously — and with little risk of downtime.

9. **Anonymity.** Finally, though we would like to believe that the human in the machine is anonymous, reality differs. It is estimated that 50% of workers have been linked to public Amazon user profiles (Lease, Hullman, & Bigham, 2013). This means that Amazon worker Ids are essentially PII. And, in fact, many workers are not blind to this.

10. **The "Superturker".** Legal scholar Dan Kahan warns particularly of a side-effect of prior, repeated exposure to cognitive studies on AMT (Kahan, 2013). Chandler, Mueller, and Paolacci (2013) note that while the probability that any one worker has seen some manipulation, there is a population of "superturkers" (prolific workers) who are significantly more likely to end up in studies. Pooling 16,408 HITs in 132 unique studies, they found that HITs were completed by 7498 unique workers. The top 10% of prolific workers completed 41% of total HITs. On a positive note, Chandler et al. (2013) note that these turkers are less likely to be multi-tasking and more likely to be available for follow-up studies such as required for longitudinal research. On the negative side, these workers are much more likely to have participated in cognitive tasks potentially biasing them for future cognitive tasks (Chandler et al., 2013).

4.4 GENERAL PROCEDURE

Each experiment in this dissertation was delivered over the Internet in a web browser. Subjects (AMT workers) were randomly assigned to groups per study design. Experiments

were implemented within a Qualtrics (http://qualtrics.com) survey using supplementary, custom JavaScript/HTML/CSS code that I wrote, as needed. Surveys were accessible via an HTTPS encrypted iframe on the AMT website (https://www.mturk.com).

In order to participate, AMT workers selected my HIT from a worker queue (Figure 2). Three base qualifications were stipulated:

- workers must be 18 years or older (provided by the Amazon Mechanical Turk Terms of Service in Appendix B);
- workers must be physically located in the United States; and,
- Workers must have a HIT success rate of 90% or better.

The AMT API provided for qualification criteria such as geographic region and HIT success rate. By specifying these as requirements, Amazon automatically matched eligible workers to my "surveys". A sample assignment (Experiment 1A) is provided in Appendix C.

Using AMT command line tools, I tested each experiment in the AMT sandbox before deploying to AMT. Requesters can monitor assignment progress, approve or reject workers, download results, and assign bonuses from their dashboard. I tested assignments in several browsers both on on a Mac and Windows PC to ensure functionality and appearance were consistent. Following successful tests, I pushed assignments to AMT.

Workers who selected one of my HITS were told they would be participating in either a language study, user interface study, image annotation study, or ethics survey, depending on the experiment. Because these were all studies of language use, workers were expected to have basic *linguistic competence* in English. It was assumed that if they were on the Internet capable of responding to a request to participate in an online experiment, and residing in the United States, they had adequate competence in English. However, I also provided a demographic survey question targeting fluency level, if the subject was not a native speaker of English.

Each experimental design was deployed as a single Qualtrics survey. Qualtrics provides for sophisticated functionality such as survey flow, block randomization, display and

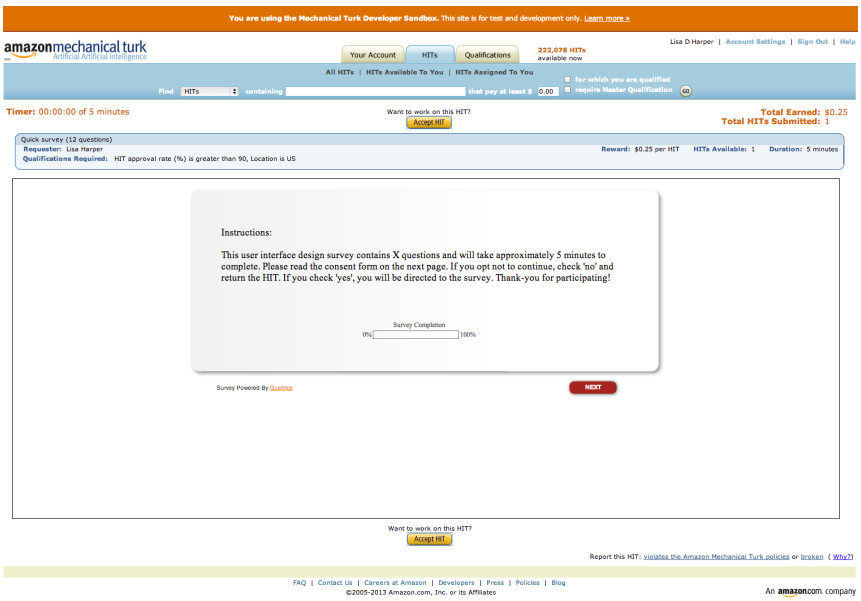

Figure 3: Participant Instructions (Experiment Three)

question logic, and a JavaScript application programming interface (API). Several of my surveys required hand-coded instrumentation not provided by Qualtrics. I discuss instrumentation in the methods section of each experiment described.

Because I conducted multiple, related experiments simultaneously, exposure to one experiment potentially disqualified a worker from other experiments. For example, Experiment 1A has a potential priming effect on the topic of Internet privacy that could potentially bias that same worker in Experiment 1B. In such cases, I included a message on the preview that said something like, "please don't accept this HIT if you have previously participated in an experiment with a blue rectangle." As reinforcement, I checked each worker on acceptance of consent against a database of prior workers. If a worker had already participated in a closely related experiment, a message was presented to that worker to please return the HIT. Then they were prohibited from further access. Eve so, it was still possible for a single worker to participate in more than one experiment. For example, if a worker signed up for Experiment 1B, they were allowed to also sign up for 2B.

In each experiment, workers previewed basic instructions (Figure 3) before selecting the HIT. I inserted custom code to prevent participants from advancing beyond this screen until the HIT had been accepted. In order to proceed, the worker must have clicked "Accept HIT".

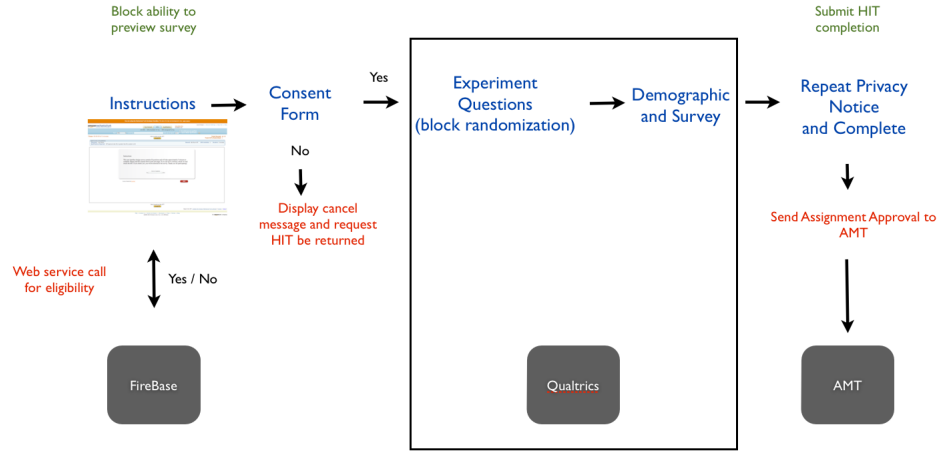

Figure 4: Experiment Flow

After clicking "next", a consent form was displayed (see Appendix A). Again, in order to continue, the participant must have selected "yes" to continue.

General experiment flow is depicted below (Figure 4).

At the completion of each survey, I configured assignments to automatically approve HITS so that Amazon would pay the worker immediately. Experiments were designed to take approximately three minutes to complete. I offered 15 cents for most. While it was also possible to manually review results before approving HITs, I decided this was not necessary for any of the studies included here since workers were recruited from those with known high performance. I did, however, manually assign random dollar bonuses after assignment completion.

4.5 DATA STORAGE AND PRIVACY

Data was stored in three places (Table 2): AMT, Qualtrics, and FireBase. AMT stored a list of workerIds for each experiment, Qualtrics stored survey data (but no workerIds), and FireBase was used to store a list of worker hashes (but no data or workerIds).

There was no connection between AMT and Qualtrics except for URL-encoded data passed to and from the Qualtrics server:

- On acceptance of a HIT, Amazon passed Qualtrics an AssignmentId, HitId, and WorkerId.

Table 2: Data Captured

Stored	Type of Data
Amazon Server	WorkerId, AssignmentId, HitId, Accept / Reject, Fees paid
Qualtrics Server	Random number identifier, Experiment data
FireBase Server	Hash of workerIds for each experiment

- On completion of the HIT, Qualtrics sent back to Amazon the AssignmentId.

Survey data was not stored on Amazon's servers nor was participant identity accessible to Qualtrics via the AMT API. Because AMT does not offer "verified" user profiles, users can lie about their demographic group to qualify for a study. However, given the relative anonymity of the system, most appear to be honest (Ipeirotis, 2010a).

The Firebase service listed above was required only to disqualify workers from participating in certain experiment combinations. To safeguard identity, my code on Qualtrics converted the AMT WorkerId to a 32-bit hash and passed this hash to the Firebase service. New prospective workerIds were hashed and compared with the hash list on the Firebase service. Hashing is uni-directional, thus it is not possible to recover the original WorkerId from a hash list alone.

On Qualtrics, the WorkerId was discarded and replaced with a random number identifier. WorkerId was treated as Personally Identifiable Information (PII) and de-linked from the data such that it would not be possible to re-link workerIDs to experiment results. Furthermore, though Qualtrics retains IP address data by default, I turned this, and other browser identifiers, off for these experiments. Because no identifying questions were asked, nor was IP address collected, all experiment data collected is fully anonymized.

The next section addresses characteristics of the population sampled.

4.6 POPULATION SAMPLE

The first and third columns of the table in Appendix D summarize questionnaire responses across the five studies described in Chapters 5 - 7. Of the 1158 subjects who took the demographic survey, 84% were unique, with 16% duplication due to participation in more than one experiment. This was possible, depending on the order in which workers accepted tasks. For example, if a worker accepted a HIT in Experiment 3, that same worker was later still eligible for any of the other experiments. However, once participating in Experiment 1B, he or she would have been no longer eligible to participate further.

Demographics are comparable to those published in other studies (Ipeirotis, 2010b; Mason & Suri, 2011; Ross, Irani, Silberman, Zaldivar, & Tomlinson, 2010), though the sample I collected was skewed more toward males. Of 1158 subjects, 73% were under 35, 53% male, 73% caucasian, 97% English speaking, 86% having at least some college, and 53% making under 30,000 per year. The profile drawn here is from a set of workers with a 90% or better HIT acceptance rate and from the United States.

In September of 2011, McDonald and Peha (2011) conducted a user survey on AMT to study the differences between what users expect "Do Not Track" to mean versus DNT definitions under debate. At the time, DNT was very new. They studied 304 participants limiting participation to the United States. 81% had never heard of DNT.

Though I did not use those exact questions in my survey, in October of 2013, 91% reportedly knew what a tracking cookie is. 73% admitted to using browser plugins for privacy protection and more than half to configuring their browsers to "opt-out". However, despite that AMT workers are much more knowledgeable than two years ago, most still do not realize that DNT, as specified by advertisers, is limited to not showing targeted advertisements: information is still collected, stored, and used.

In an older study (not on AMT), Acquisti and Grossklags (2005) found a correlation between concern for privacy and income. Those with higher incomes were generally more concerned with privacy. However, this may have correlated more strongly with knowledge than income. They noted that their

sample particularly lacked knowledge about technological or legal forms of privacy protection.

4.7 VALIDITY

Discussed below are three aspects of validity for studies using Mechanical Turk: internal, external, and construct validity. Questions of statistical validity are addressed in the results sections of experiments.

4.7.1 *Internal Validity*

Internal validity concerns the equivalence between groups and the control of extraneous variables. Other behavioral experiments utilizing AMT (Crump, McDonnell, & Gureckis, 2013) have suggested that the pool of available workers is sufficiently representative and large such that random assignments consistently produce results equivalent to those produced in carefully controlled laboratory settings. Furthermore, it is possible to conduct experiments where workers may are not aware that they are in an experiment (one source of experimental bias). Also, Paolacci et al. (Paolacci et al., 2010) found that retention was particularly high on AMT, compared to other pools compared (university student population and Internet boards). However, they noted that, unlike student populations, turker membership is organic and workers may be potentially available for years. This particularly highlights the need to track responses across experiments.

More insidious to internal validity may be the effects of payment itself. Two effects of *volunteer bias* are of particular concern are: 1) effect of informed consent; and 2) effects of obligation incurred by accepting a HIT and, as as result, payment for services. Rush, Phillips, and Panek (1978) found difference between unpaid volunteers and paid subjects in a selective attention task. Unpaid volunteers were found to commit fewer omission errors than paid subjects.

Another potential threat to internal validity is the control of the subject environment. Mentioned previously, was the problem of ensuring all conditions are presented consistently across browser types and monitor sizes. In addition, there is no way to control what sorts of environmental factors may be present: workers may be participating in multiple tasks, watching television, or be distracted in a myriad of ways.

To address these concerns, I took the following precautions:

1. using custom code, I limited participation across conditions where prior exposure to privacy questions was potentially biasing; and,
2. to the extent possible, I sandbox tested — not just code consistency, but visual consistency across platforms and browsers.

4.7.2 *External Validity*

One important aspect of external validity is whether AMT workers are representative of the desired population as a whole. The experiments described in this dissertation rely on basic linguistic competency. For this purpose, native language competency is of concern. Mechanical Turk appears representative of the U.S. population as a subject pool from the perspective of gender, race, and education. Previous studies (Behrend et al., 2011; Buhrmester, Kwang, & Gosling, 2010) find that turkers are slightly more representative of U.S. population statistics than are standard Internet samples. They are also significantly more diverse that university samples.

The population sample participating in studies described here may, however, differ from the general Internet population in other ways. Turkers polled seemed surprisingly knowledgeable of privacy issues. This indicates that experimental results, in some cases, may not be representative of a more general population.

4.7.3 *Construct Validity*

To large extent, construct validity is a property of each individual experiment. However, the experiments described in this dissertation fall into the genre of both linguistic and judgment /decision-making tests. To this end, there exist other studies which are generally comparable.

Paolacci et al. (2010) conducted replication of traditional judgment and decision-making tasks on AMT and compared them with groups from traditional subject pools at a large University and also Internet discussion boards. Tests included the Kahneman and Tversky (1984) Asian disease problem, Kahneman and Tversky (1983) Linda problem, and Baron

and Hershey (1988) Physician's problem. Their results confirm that AMT is a reliable source of experimental data in judgment and decision-making. Not only were group results similar across conditions, but the effect size for AMT was the highest. More generally, AMT has been validated as a tool for behavioral cognitive studies such as those in reaction time research (e.g., Stroop, Task-switching, Flanker tasks, etc.; (Crump et al., 2013) and recall of written information (Tietze, Winterboer, & Moore, 2009).

While researchers in natural language processing were among the first to utilize AMT for the collection and annotation of linguistic resources (Callison-Burch & Dredze, 2010), its adoption by theoretical linguists has been slow to develop. Traditionally, user judgments supporting theoretical claims have been weakly quantitative relying on a very small number of examples and researchers – often limited to the authors of theoretical studies themselves. Gibson and Piantadosi (2011) demonstrate the utility of AMT for collecting linguistic behavioral judgments for acceptability of sentence / meaning pairs. Sprouse (2010) compares just such a task with two groups of users (AMT versus laboratory) each with 176 participants. Data collected from the two groups was deemed virtually indistinguishable.

Finally, researchers in pragmatics (e.g., in studies of implicature) have begun not only to adopt experimental methods, but have applied these to collections on AMT (for example, Bergen, Goodman, & Levy, 2012; Degen & Tanenhaus, 2011; Stiller, Goodman, & Frank, 2011). As noted by Anand and Andrews (2011), pragmatic inference depends on a multitude of factors including task structure, social norms, and response elicited. Furthermore, because there are potentially so many parameters, it is difficult to systematically model interactions between linguistic form, context, and pragmatic inference. Crowdsourcing platforms make such study more tractable, though we have much yet to learn about what specific methods are most amenable. Problematic for pragmatic studies, in particular, is that the subject's knowledge of the experiment itself plays directly into context (Rosnow & Rosenthal, 1976).

4.8 SUMMARY

This chapter outlined a research program with the substantive hypothesis that some user confusion stemming from online behavioral advertising is caused by error in discourse understanding. I described three situations where miscommunication may occur. The first considers the possibility of pragmatic implicature in a "do not track" modal dialog, the second considers the indexicality of an icon embedded in an image-based advertisement, and the third the effect of visual presence on non-ratified participants on a web page.

In this chapter, I also discussed related research utilizing Amazon Mechanical Turk, outlined general procedures and considerations, and described general characteristics of the population sampled.

In each of the next three chapters, after framing experimental hypotheses within the theoretical framework, I detail specifics in terms of instrumentation, collection, analysis and results.

A

PARTICIPANT CONSENT FORM

I. INTRODUCTION/PURPOSE:
I am being asked to participate in a research study.The purpose is to research a topic in user interface design. My involvement in this study will begin when I agree to participate and will continue until the survey has completed.

II. PROCEDURES:
As a participant in this study, I will be asked to complete an online survey launch to an external website from Amazon Mechanical Turk (AMT). My participation in this study is voluntary and I am free to withdraw or discontinue participation at any time.

My participation in this study will last for approximately 3 minutes.

III. RISKS AND BENEFITS:
IV. My participation in this study may cause some discomfiture. I can stop at any point. If I complete the survey, I will receive monetary compensation for my time.

V. CONFIDENTIALITY:
My data in this study is anonymized. I will NOT be identified by AMT workerID. Amazon will not have access to survey data stored on Qualtrics servers. Qualtrics will not have access to my workerId nor IP address. No personally identifiable data is collected or stored. By signing this form, I allow the research study investigator to make these records available to the University of Baltimore Institutional Review Board (IRB) and regulatory agencies as required to do so by law. Fully anonymized data may be made available to other researchers after the study is complete.

VI. SPONSOR OF THE RESEARCH:
This research study is for a doctoral dissertation.

VII. COMPENSATION/COSTS:
I will be paid 15 cents for my participation.

VIII. CONTACTS AND QUESTIONS:
The principal investigator(s), Lisa Harper and Dr. Kathryn Summers has offered to and has answered any and all questions regarding my participation in this research study. If I have any further questions, I can contact them at lisa.harper@ubalt.edu. For questions about rights as a participant in this research study, contact the UB IRB Chair: Eric Easton, Chair, University of Baltimore Institutional Review Board, 410-837-4874, eeaston@ubalt.edu.

Certification:

I have read and understood the above information:

- ○ Yes
- ○ No

>>

B

AMT TERMS OF SERVICE

8/16/13 Amazon Mechanical Turk - Participation Agreement

amazonmechanical turk Artificial Artificial Intelligence

Lisa D Harper | Account Settings | Sign Out | Help

Your Account | HITs | Qualifications | 185,183 HITs available now

Find HITs containing that pay at least $ 0.00 for which you are qualified require Master Qualification GO

Amazon Mechanical Turk Participation Agreement

Last updated: November 1, 2012

Welcome to the Amazon Mechanical Turk services platform.

BY REGISTERING FOR AND USING THE SITE, YOU CERTIFY THAT (1) YOU ARE AT LEAST 18 YEARS OLD; (2) YOU HAVE THE AUTHORITY TO ENTER INTO THIS AGREEMENT AND BIND YOURSELF OR THE COMPANY YOU REPRESENT; (3) YOU AUTHORIZE THE ELECTRONIC TRANSFER OF FUNDS TO YOUR BANK ACCOUNT IN ACCORDANCE WITH SECTION 4 OF THIS PARTICIPATION AGREEMENT; AND (4) YOU AGREE TO BE BOUND BY ALL TERMS AND CONDITIONS OF THIS AGREEMENT, INCLUDING THE TERMS AND CONDITIONS OF THE PAYMENT SERVICE DESCRIBED IN SECTION 4 AND ALL APPLICABLE POLICIES, PROCEDURES AND GUIDELINES. This Participation Agreement (the "**Agreement**") is between you and Amazon Mechanical Turk (as defined below) and governs your and Amazon Mechanical Turk's respective rights and obligations with respect to your offering for sale, selling, requesting, purchasing, and/or providing Services (defined below) on or through the Site (as defined below).

For purposes of this Agreement, (a) "**Amazon Mechanical Turk**", "**we**", "**us**" or "**our**" means Amazon Mechanical Turk, Inc. a Delaware Corporation, (b) "**Site**" means the Amazon Mechanical Turk web site located at mturk.amazon.com, requester.mturk.com, www.mturk.com and any successor website thereto, including all services provided by us to you through the service platform on the Site, (c) "**Services**" means any service that you sell, offer to sell, request, purchase, and/or provide on or through the Site, (d) "**Affiliate**" means any entity controlled by, in control of, or under common control with Amazon Mechanical Turk, (e) "**Requester**" means you, if you use the Site to request that a Provider perform Services, (f) "**Provider**" means you, if you use the Site to perform Services for a Requester, (g) "**Amazon Account**" means any customer account that you have established with a website owned or controlled by Amazon or its Affiliates, or operated by Amazon or its Affiliates on behalf of third parties, including without limitation those websites currently located at http://www.amazon.com, http://www.amazon.co.uk, http://www.amazon.de, http://www.amazon.fr, http://www.amazon.ca, http://www.amazon.co.jp and http://www.joyo.com, and any successor or replacement websites.

This Agreement consists of the terms and conditions set forth in this document together with all applicable policies, procedures and/or guidelines that appear on the Site from time to time (collectively, the "**Policies**" which are hereby incorporated by this reference into, and made part of, this Agreement). Amazon Mechanical Turk reserves the right to change any of the terms and conditions contained in this Agreement and/or any Policies governing the Site, at any time, in its sole discretion. Any changes will be effective upon posting of the Agreement or Policies on the Site and may be made without any other notice of any kind. You are at all times responsible for reading and understanding each version of this Agreement and the Policies. YOUR CONTINUED USE OF THE SITE FOLLOWING AMAZON MECHANICAL TURK'S POSTING OF ANY CHANGES WILL CONSTITUTE YOUR ACCEPTANCE OF SUCH CHANGES. IF YOU DO NOT AGREE TO ANY CHANGES TO THIS AGREEMENT (INCLUDING TO ANY OF THE POLICIES INCORPORATED HEREIN), DO NOT CONTINUE TO USE THE SITE.

1. **Registration.**

   a. **Registration**. When you register with the Site, you will be asked to provide us with, at a minimum, your name, a valid email address, your phone number, and your physical address. Providers may also be asked to provide certain tax information at registration or afterwards. You agree to provide us with true and accurate information, and to update that information to the extent it changes in any way. When registering or updating your information, you will not impersonate any person or use a name that you are not legally authorized to use.

   You may register with the Site either by (i) using your existing Amazon Account or (ii) creating a new Amazon Account. If you do not have an existing Amazon Account at the time you register with the Site, an Amazon Account on the Amazon.com website located at http://www.amazon.com (hereinafter, "**Amazon.com**") will be automatically and concurrently established in your name with the same e-mail address and password you provide to us. Amazon Accounts used in conjunction with the Site are governed by the Conditions of Use and Privacy Notice applicable to Amazon.com, as well as the Amazon Mechanical Turk Privacy Notice. You may not use multiple Amazon Accounts to register with Mechanical Turk. Your Amazon Account username must not suggest affiliation with Amazon, Amazon Mechanical Turk, or any third party unless that third party specifically gave you permission to do so.

   b. **Passwords and Account Use**. You are solely responsible for maintaining the secrecy and security of your password. You may not disclose your password to any third party (other than third parties authorized by you to use your account) and are solely responsible for any use of or action taken under your password on the Site. If your password is compromised, you must change your password. You may not permit any other person to perform Services as Provider using your Amazon Account. Additionally, if you are using the Site as a Provider, you may not use different Amazon Accounts to perform Services.

2. **Amazon Mechanical Turk's Role**. Amazon Mechanical Turk provides a venue for third-party Requesters and third-party Providers to enter into and complete transactions. Amazon Mechanical Turk and its Affiliates are not involved in the transactions between Requesters and Providers. As a result, we have no control over the quality, safety or legality of the Services, the ability of Providers to provide the Services to Requesters' satisfaction, or the ability of Requesters to pay for Services. We are not responsible for the actions of any Requester or Provider. We do not conduct any screening or other verification with respect to Requesters or Providers, nor do we provide any recommendations. As a Requester or a Provider, you use the Site at your own risk.

3. **Your Use of the Site**

   a. **Requesters in General**. Upon completion of Services to Requesters' reasonable satisfaction, Requesters must pay Providers for their Services. As a Requester, you agree that upon your approval of the Services performed by a Provider, payment will be remitted to the Provider automatically (as described in Section 4 below). After you have approved the applicable Services, you are not entitled to any refund of your payment for such Services. If a Requester is not reasonably satisfied with the Services, the Requester may reject the Services. As a Requester, you will be charged a fee for your use of Amazon Mechanical Turk in connection with each request for Services. Please review the applicable Amazon Mechanical Turk Fees contained in the Policies for all applicable fees associated with your use of the Site pursuant to this Agreement. All fees are in U.S. dollars unless stated otherwise. The Amazon Mechanical Turk Fees may vary in the future. You agree to pay the amounts set forth in the Amazon Mechanical Turk Fees from time to time on the terms set forth herein and therein, and to check the fees and terms each time you use the Site. You acknowledge that, while Providers are agreeing to perform Services for you as independent contractors and not employees, repeated and frequent performance of Services by the same Provider on your behalf could result in reclassification of that employment status. If you have any questions about your obligations to comply with local laws and regulations pursuant to Section 6, you should seek independent legal advice. To the extent you receive any contact or personal information regarding any Provider who has performed Services for you, such information may only be used as necessary for you to comply with applicable laws and for no other purpose whatsoever. Further, you agree that you will only accept work product from Providers that has been submitted through the Site.

   b. **Providers in General**. You may only register once with Mechanical Turk as a Provider. Providers may perform Services for any Requester in accordance with the specifications submitted by the Requester. However, if the Services do not meet the Requester's reasonable satisfaction, the Requester may reject the Services and repost the specific request. As a Provider, the Requester for whom you provide Services is your client, and as

https://www.mturk.com/mturk/conditionsofuse 1/4

C

SAMPLE AMT ASSIGNMENT DEFINITION

```
title = Language study and demographic survey (3 min)
description: one task and then just a survey
keywords:survey, language

# how much you'll pay each subject
reward:.15

# how many subjects do you want
assignments:30

#######################################
## HIT Timing Properties
#######################################

# 60*10, 10 mins to finish a suvey
assignmentduration:600

# 60*60*24*2, 2 day to keep on mturk
hitlifetime:172800

# 10 seconds to auto approve the response
#
autoapprovaldelay:10

#######################################
## Qualification Properties
#######################################

# user must have an approval rate of 90% or greater
qualification.1:000000000000000000L0
qualification.comparator.1:greaterthan
qualification.value.1:90
qualification.private.1:false

# user must be in the United States
qualification.2:00000000000000000071
qualification.comparator.2:equalto
qualification.locale.2:US
qualification.private.2:true
```

D

SURVEY RESPONSE

Table 3: Aggregate Demographic Profile and Survey Response

Category	Experiment 3	Total
Age		
18-24	29%	28%
25-34	44%	45%
35-49	17%	20%
50-64	10%	7%
64+	0%	0%
Gender		
Male	63%	53%
Female	37%	47%
Ethnic Identity		
American Indian / Native American	0%	1%
Asian or Pacific Islander	10%	10%
Black / African American	3%	7%
Hispanic or Latin American	5%	7%
White / Caucasian	82%	73%
Near Eastern or Arabic	0%	0%
Other	2%	2%
Native Language		
English	99%	97%
Other	1%	3%
English		
A little English	0%	0%
Some English	0%	0%

Fluent English	0%	60%
Near-native English	100%	40%
Education		
Some High School	1%	1%
High School Graduate	17%	13%
Some College or Associate Degree	39%	40%
College Degree	35%	37%
Post-graduate Degree	8%	9%
Income		
Under 20,000	30%	34%
20,000 - 30,000	25%	19%
30,000 - 40,000	13%	15%
40,000 - 50,000	13%	10%
50,000+	20%	22%
IT Job		
Yes	21%	20%
No	79%	80%
Internet Usage		
Fewer than 4 hours per week	1%	2%
4-10 hours per week	13%	11%
10-25 hours per week	41%	30%
25+ hours per week	45%	57%
Shop Online		
Never	1%	1%
Rarely	20%	17%
Sometimes	59%	30%
Often	21%	52%
Sense of Privacy in Public		
Not private	31%	29%
A Little private	34%	30%
Somewhat private	23%	29%

Private	11%	9%
Very prvate	1%	3%
Sense of Privacy at Home		
Not private	5%	4%
A Little private	15%	10%
Somewhat private	23%	23%
Private	34%	37%
Very prvate	24%	26%
Importance of Online Privacy		
Not much of an issue	8%	7%
Somewhat important	50%	51%
Really important	42%	42%
Steps to Protect Privacy		
Don't know how to protect	12%	8%
Know how but not consistent	38%	49%
Know how and take measures	50%	43%
Know what a tracking cookie is		
Yes	90%	91%
No	10%	9%
DNT means...		
Do not show targeted advertising	17%	20%
Do not trakc across sites	44%	54%
Do not track on this site	49%	47%
Do not collect information	39%	48%
Do not store information	38%	43%

Would Turn off Tracking if Easy?		
Yes	88%	91%
No	12%	9%
Browser Configured Opt-Out		
Yes	54%	55%
No	46%	45%
Use a Browser Plugin for Privacy Protection		
Yes	70%	73%
No	30%	27%
Total	**100**	**1158**

E

PARTICIPANT SURVEY

Q	Type	Question	Answer Type	Answer 1	Answer 2	Answer 3	Answer 4	Answer 5
1	Demographic	How old are you?	Radio	18-24	25-34	35-49	50-64	65+
2	Demographic	What is your gender?	Radio	Female	Male			
3	Demographic	What is your ethnic identity?	Radio	American Indian / Native American	Asian or Pacific Islander	Black / African American	Hispanic / Latino	White / Caucasian
4	Demographic	What is your native language?	Radio	English	Other			
5	Demographic	What is your English proficiency?	Radio	A little English	Some English	Fluent English	Near-native English	
6	Demographic	What is your education?	Radio	Some high school	High school graduate	Some college / Associate's degree	College graduate	Post-graduate degree
7	Demographic	What is your income level?	Radio	Under 20,000	20,000-30,000	30,000-40,000	40,000-50,000	50,000+
8	Technical Savvy	Do you either have a job in an IT-related field (for example, computer science, web development or similar; are a student in an IT-related area?	Radio	Yes	No			
9	Technical Savvy	What is your Internet usage?	Radio	fewer than 4	4 to 10	10 to 25	25+	
10	Technical Savvy	How much do you shop online?	Radio	Never	Occasionally	Often	Most of my shopping is online	
11	Privacy	How private do you feel on the Internet when you are using your computer (i.e., laptop, tablet) in public (i.e., a coffee shop, library, school, etc.)	Radio	Not private	A little private	Somewhat private	Private	Very private
12	Privacy	How private do you feel on the Internet when you are using your computer (i.e., laptop, tablet) at home?	Radio	Not private	A little private	Somewhat private	Private	Very private
13	Privacy	How important is online privacy to you?	Radio	Not much of an issue / I hardly think about it	A somewhat important issue that I think about sometimes	A really important issue I think about a lot		
14	Privacy	What steps do you take to protect your privacy?	Radio	I don't really know how to protect my personal information online	I know how to protect my personal information online but don't consistently do so	I know how to protect my privacy online and consistently take the necessary steps to do so		
15	Privacy	Do you know what web browser tracking cookies are?	Radio	No	Yes			
16	Privacy	Do you know what "do not track" means?	Multiple Select	Do not show targeted advertising	Do not track me across sites	Do not track me on this site	Do not collect information on me	Do not store information on me
17	Privacy	If you could easily turn off online tracking, would you?	Radio	No	Yes			
18	Privacy	Do you opt-out of online tracking in your browser settings?	Radio	No	Yes			
19	Privacy	Do you use any browser plug-ins such as adblockers or other privacy enhancing tools on your computer?	Radio	No	Yes			

BIBLIOGRAPHY

Acquisti, A., & Grossklags, J. (2005). Privacy and rationality in individual decision making. *IEEE Security & Privacy*, 24–30.

Acquisti, A., John, L. K., & Loewenstein, G. (2012). The impact of relative standards on the propensity to disclose. *Journal of Marketing Research*, *49*(2), 160–174.

Alonso, O., & Mizzaro, S. (2009). Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, 15–16.

Ambati, V., & Vogel, S. (2010). Can crowds build parallel corpora for machine translation systems? In *CSLDAMT '10 Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (pp. 62–65).

Anand, P., & Andrews, C. (2011, July 27). Assessing the pragmatics of experiments with crowdsourcing: The case of scalar implicatures [slides]. In *Crowdsourcing Technologies for Language and Cognition Studies.* Boulder. Retrieved from http://www.crowdsourcing.org/document/assessing-the-pragmatics-of-experiments-with-crowdsourcing-the-case-of-scalar-implicatures/5558

Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology*, *54*, 569-579.

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods*, *43*(3), 800–813.

Bergen, L., Goodman, N. D., & Levy, R. (2012). That's what she (could have) said: How alternative utterances affect language use. In *Cogsci 2012* (pp. 1–6). Sapporo.

Buhrmester, M., Kwang, T., & Gosling, S. (2010). Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 1–23.

Callison-Burch, C., & Dredze, M. (2010). Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating*

Speech and Language Data with Amazon's Mechanical Turk (pp. 1–12).

Chandler, J., Mueller, P., & Paolacci, G. (2013, July). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Accepted for publication at Behavior Research Methods*.

Chiarella, S. (2013, Feb 11). *Subject: Does mturk have a new policy excluding international workers? [email].* Tips for Requesters on Mechanical Turk. Retrieved from http://turkrequesters.blogspot.com/2013/01/the-reasons-why-amazon-mechanical-turk.html

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013, March). Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PLoS ONE*, *8*(3), e57410.

Degen, J., & Tanenhaus, M. K. (2011, June 2 – 4). A constraint-based approach to scalar implicature processing. In *Poster presented at experimental pragmatics 2011* (Vol. 51, pp. 437–457).

Gibson, E., & Piantadosi, S. (2011). Using mechanical turk to obtain and analyze english acceptability judgments. *Language and Linguistics Compass*, *5*, 509 – 524.

Gormley, M. R., Gerber, A., Harper, M., & Dredze, M. (2010, June). Non-expert correction of automatically generated relation annotations. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (pp. 204–207). Los Angeles: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/W10-0732

Hastak, M., & Culnan, M. J. (2010, January). *Future of privacy forum: Online behavioral advertising "icon" study* (Tech. Rep.).

Howe, J. (2006). The rise of crowdsourcing. *Wired Magazine*.

Ipeirotis, P. (2010a). Analyzing the Amazon Mechanical Turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students*, *17*(2).

Ipeirotis, P. (2010b). Demographics of Mechanical Turk. *NYU Working Paper*.

Kahan, D. (2013, 10). Fooled twice, shame on who? problems with mechanical turk samples. *Cultural Cognition Project at Yale Law School*. Retrieved from http://www.culturalcognition.net/blog/2013/7/10/fooled-twice-shame-on-who-problems-with

-mechanical-turk-stud.html

Kahneman, D., & Tversky, A. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgement. *Psychological Review*, *90*.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, *39*(4), 1–11.

Komanduri, S., Shay, R., Norcie, G., & Ur, B. (2012). AdChoices? Compliance with online behavioral advertising notice and choice requirements. *CMU-CyLab-11-005*, 1–22.

Lanier, J. (2006). Digital maoism. *The Edge*, *183*.

Lanier, J. (2010). *You Are Not a Gadget*. Random House LLC.

Lease, M., Hullman, J., & Bigham, J. (2013). Mechanical Turk is not anonymous. *Social Science Research Network*.

Leon, P. G., Ur, B., Shay, R., Wang, Y., Balebako, R., & Cranor, L. (2012). *Why Johnny can't opt out: A usability evaluation of tools to limit online behavioral advertising* (Tech. Rep. No. CMU-CyLab-11-017).

Mason, W., & Suri, S. (2011, June). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*(1), 1–23.

Mason, W., & Watts, D. J. (2009, June). Financial incentives and the "performance of crowds". In *HCOMP '09: Proceedings of the ACM SIGKDD Workshop on Human Computation.* ACM.

Mayer, J., & Mitchell, J. C. (2012, May). Third-party web tracking: Policy and technology. In *SP '12: Proceedings of the 2012 IEEE Symposium on Security and Privacy.* IEEE Computer Society.

McDonald, A. M., & Peha, J. (2011). Track gap: Policy implications of user expectations for the 'do not track' internet privacy feature. *TPRC*.

Munro, R., & Tily, H. (2011, August). The start of the art: An introduction to crowdsourcing technologies for language and cognition studies. In *Workshop in Crowdsourcing Technologies for Language and Cognition Studies* (pp. 1–10). Boulder.

Nowak, S., & Rüger, S. (2010). How reliable are annotations via crowdsourcing. In *MIR '10 Proceedings of the International Conference on Multimedia Information Retrieval* (pp. 557 – 566). New York: ACM.

Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running experiments on amazon mechanical turk. *Judgment and*

Decision Making, *5*(5), 411–419.

Reips, U.-D. (2006). Standards for Internet-based experimenting. *Experimental Psychology (formerly Zeitschrift für Experimentelle Psychologie)*, *49*(4), 243–256.

Rosnow, R. L., & Rosenthal, R. (1976). The volunteer subject revisited. *Australian Journal of Psychology*, *28*, 97-108.

Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., & Tomlinson, B. (2010, April). Who are the crowdworkers?: Shifting demographics in mechanical turk. *CHI EA '10: CHI '10 Extended Abstracts on Human Factors in Computing Systems*.

Rush, M. C., Phillips, J. S., & Panek, P. E. (1978). Subject recruitment bias: The paid volunteer subject. *Perceptual and Motor Skills*, *47*(2), 443–449.

Saffer, D. (2013). *Microinteractions*. O'Reilly Media.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. , 254–263.

Sprouse, J. (2010). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, *43*(1), 155–167.

Stiller, A., Goodman, N. D., & Frank, M. C. (2011, Jul). Ad-hoc scalar implicature in adults and children. In *Proceedings of the 33rd annual meeting of the cognitive science society.*

Tietze, M. I., Winterboer, A., & Moore, J. (2009, March). The effect of linguistic devices in information presentation messages on comprehension and recall. In *Proceedings of the 12th European Workshop on Natural Language Generation. Association for Computational Linguistics* (pp. 114–117).

Tribal Fusion. (2012). *AdChoices Our experience.* Retrieved from http://www.iabuk.net/sites/default/files/Tribal%20Fusion.pdf

Ur, B., Leon, P. G., Cranor, L. F., Shay, R., & Wang, Y. (2012). *Smart, useful, scary, creepy: perceptions of online behavioral advertising* (Tech. Rep. No. CMU-CyLab-12-007).