Is Your Inseam a Biometric? A Case Study on the Role of Usability Studies in Developing Public Policy

Rebecca Balebako, Richard Shay, Lorrie Faith Cranor
Carnegie Mellon University
Pittsburgh, PA
(balebako, rshay, lorrie)@cmu.edu

***Abstract*—In this paper, we present a case study of applying usable privacy methodologies to inform debate regarding a multi-stakeholder public policy decision. In particular, the National Telecommunications and Information Administration (NTIA) relied on a multi-stakeholder process to define a set of categories for short-form privacy notices on mobile devices. These notices are intended for use in a United States national code of conduct to assist mobile device users in making decisions regarding data collection. We describe, specifically, a 791-participant online study to determine whether users consistently understand these proposed categories and their definitions. We found that many users did not understand the terms in our usability study. The heart of our contribution, however, is a case study of our participation in this group as academic usable privacy and security experts, and a presentation of lessons learned regarding the application of usable privacy and security methodology to public policy discussion. We believe this work is valuable to usable privacy and security researchers wishing to affect public policy.**

I. Introduction

Public policy and regulation intersect with human computer interaction in many domains. In areas including voting machines, accessibility, and privacy, regulators may try to step in where market forces have failed. We argue that policy-making can and should be be informed by usability studies, and those policies in these areas that are not informed by the usable privacy and security community may be ineffective. We provide a case study of a multi-stakeholder process to standardize smartphone privacy notices in the United States. We present a user study, which we ran near the end of the process, that demonstrates the shortcomings of failing to take usability into account throughout the process. Furthermore, we discuss what lessons can be learned from our experience.

The U.S. National Telecommunications and Information Administration (NTIA) initiated a multi-stakeholder effort to develop a standardized short-form privacy disclosure on mobile devices. These standardized disclosures will show the user both what data is being shared and with which entities it is being shared. The year-long NTIA multi-stakeholder process (NTIA MSHP) lasted from June, 2012 to July, 2013. The NTIA stakeholders – representing app developers, consumer groups, and government – developed a Code of Conduct that provides guidance for app developers for a short-form privacy notice. This Code outlines seven categories of data and eight categories of third-party entities that apps should include in short-form privacy notices, but does not specify a format for these notices.

We present the results of a 791-participant online study in which we investigate whether participants are able to categorize realistic data-sharing scenarios using the NTIA MSHP categories. We also present the same categorization from four experts who participated in the NTIA MSHP process. Of the 52 examples given in our scenarios, participants showed low common agreement for how to classify the data or entity in 23 cases. Overall, we found that many of the proposed categories and definitions were not consistently understood by our participants, including our expert participants. We discuss categories that need clarification, and offer suggestions for improving the Code based on our findings. This study was undertaken by the authors independent from the NTIA MSHP, and is the first and only human-subject study conducted on those categories to date. Our results suggestion, further, that user studies should have been a larger part of the NTIA MSHP.

Based on our experience, we provide lessons learned for usable privacy and security experts who wish to participate in multi-stakeholder processes. Our contribution is primarily the insights and recommendations based on our experience participating in this process. We hope our discussion will enable and encourage usability experts to participate in public policy processes more readily and advocate for legislation and codes that are better informed by user studies.

We first discuss the background of the NTIA MSHP, along with related work. We then describe the methodology of our user study. We next present the high-level results from our user study. We explore the limitations of our study. Finally, we discuss recommendations and lessons learned for usability researchers who wish to improve policy making.

Permission to freely reproduce all or part of this paper for noncommercial purposes is granted provided that copies bear this notice and the full citation on the first page. Reproduction for commercial purposes is strictly prohibited without the prior written consent of the Internet Society, the first-named author (for reproduction of an entire paper only), and the author's employer if the paper was prepared within the scope of employment.
USEC '14, 23 February 2014, San Diego, CA, USA
Copyright 2014 Internet Society, ISBN 1-891562-37-1
http://dx.doi.org/10.14722/usec.2014.23 039

II. Usability and Public Policy

Usability and consumer testing have previously played a role in developing standards and policy for technology. We discuss several examples of usable privacy and security experts who have been involved in the public policy process. We first examine privacy issues in particular, and then briefly discuss two other examples of usability and public policy: voting machines and accessibility.

Independent academic research has evaluated privacy standard proposals created by the the The World Wide Web Consortium (W3C). Although Platform for Privacy Preferences (P3P) does not include a user interface standard, usability tests of prototype user agents conducted independently by members of the working group informed the standard's User Agent Guidelines [1]. While the Do Not Track (DNT) process refrained from defining a user interface, one independent academic user study examined the usability of an implementation of DNT [2], and another user study performed by the chair of the W3C DNT working group examined user understanding of DNT [3].

Academic researchers have also proposed privacy label standards. Kelley et al. developed and tested a "privacy nutrition label" for websites. They also found that a tabular format was liked by users and facilitated policy comparison [4].

Consultants have also been engaged by policy makers to evaluate the usability of standards. For example, The U.S. Gramm-Leach-Bliley Act of 1999 (GLBA) required financial institutions to provide a privacy notice to customers. The Federal Trade Commission hired Kleimann Communication Group to conduct user studies and develop a privacy notice prototype. Their qualitative research involved iterating over prototypes with several studies — including focus groups and usability testing [5]. This prototype was then tested against several others with quantitative testing [6], and the results were used to develop the final 'model' form, which presents information in a tabular format [7].

Think tanks and user-advocacy groups have also been engaged in evaluating the usability of privacy notices and icons. For example, the Future of Privacy Forum researched consumer's responses to notices about online behavioral advertising (OBA). They found that transparency and choice increased people's comfort with OBA. That study also compared the effectiveness of different icons in communication about OBA [8]. Unfortunately, the icon revealed as the most effective was not selected by the ad industry.

Usability issues with voting came to national attention during the 2000 U.S. presidential campaign. In 2002, Congress passed the Help America Vote Act, which included helping states evaluate their voting systems. Norden et al. discuss voting machine and ballot usability, and provide four case studies in which usability experts evaluated voting systems and worked with public officials to improve usability [9].

Another issue at the crossroads of usability and public policy is the development of accessibility standards. Lewis and Treviranus explain that public policy impacts the accessibility of information technology content and services by influencing funding, setting standards, enforcing regulation, and promoting adoption. For example, Section 508 of the Rehabilitation Act helps adoption of standards by requiring all federal websites to meet accessibility standards. The authors encourage participation in standards development or related activities to influence public policy [10].

III. Privacy and Smartphone Users

In this section, we discuss smartphone users' concerns about smartphone privacy, including which categories of data may be privacy-sensitive for users. We then discuss related work on designing usable privacy notifications.

A. User Concerns about Smartphone Privacy

Several categories of smartphone data raise privacy concerns, including the categories mentioned in the NTIA MSHP Code of Conduct. Biometric data can serve as a unique identifier for linking to a user's other activities [11]. These unique identifiers can cause particular privacy concern as they often cannot be revoked or changed, even when stolen [12]. Users' concerns about the collection of their browsing history have been documented a number of times [13]–[15]. Additional privacy issues inherent in the collection of metadata, such as logs of browsing, phone calls, or text messages, have been publicized in the wake of revelations about the U.S. National Security Agency's PRISM program. Phone usage data and metadata can be collected by apps and used to infer hobbies, medical conditions, and beliefs [16]. Users' beliefs and activities can often be inferred from the people with whom they associate [16]. The collection of users' contacts has led to privacy outrage in the past, such as when Facebook's smartphone app was discovered uploading the names and phone numbers from users' address books to Facebook's servers without providing notice [17]. The metadata from users' emails alone can be used to infer their real-life social network and associations [18], [19]. Furthermore, the fact that data is collected can have a chilling effect on individuals' free speech [20], and most individuals would likely be unaware when their data and metadata could reveal them to be violating the law [21].

Sensitive information may exacerbate privacy concerns. Financial information can cause privacy issues both because individuals might be loath to disclose information about their earnings, as well as fear about the potential of price discrimination [22]. Similarly, privacy is fundamental to a doctor-patient relationship, and disclosure of health information could cause financial harm if used by a health-insurance company to deny coverage to a patient [23].

Location data can also arouse privacy concern, particularly when the location is not visited by many people [24] or when the location information is highly granular [25]. Users also believe their files, such as photos and videos, to be sensitive [26]. Furthermore, nearly all participants in a study by Felt et al. would have been upset if the text messages and emails stored on their phone were shared publicly [27]. Information collected by fitness apps may include sensitive information that could be sold to insurance companies [28].

In addition to the type of data, users are concerned about with whom the data is shared. Social networks, government, and advertisers may all be of particular concern. A Pew Research Study found that 63% of Americans would feel their privacy had been violated if they knew the government had collected information about their calls and online communication [29]. In addition, social networks may be a concern due to the accidental leakage of private information (willingly provided by the user) to unanticipated parties [30], [31]. Urban et al. found survey participants were unwilling to share contact information with advertisers [32].

B. Usability Issues

Several studies have examined smartphone privacy notifications. An Internet survey of 308 Android users and a laboratory study of 25 Android users found that only 17% paid attention to the permissions when installing an application. They also found that only 3% of the Internet survey respondents demonstrated full comprehension of the permissions screen [33]. Kelley et al. found that when Android users were presented with privacy information, they chose apps with fewer permission requests [34]. Balebako et al. examined users' reactions to a user interface displaying information about data collected, and found users were surprised by the quantity and destinations of data. Additionally, smartphone users often did not recognize the names of third-party advertisers or data aggregators with which smartphone games shared data [35]. Felt et al. proposed a framework for smartphone platforms to request permission for data from the user [36].

IV. Multi-stakeholder processes in Privacy Policy

In this section we describe how public policy in the US has addressed mobile privacy notices.

In 2012, the White House issued a report on consumer data privacy, which included a Consumer Privacy Bill of Rights [37]. The second principle in the bill of rights is transparency, which is summarized as: "Consumers have a right to easily understandable and accessible information about privacy and security practices." The White House report emphasizes the role of multi-stakeholder processes to develop and define privacy practices and technologies, and to develop "enforceable codes of conduct." It calls upon the Department of Commerce's National Telecommunications and Information Administration (NTIA) to lead multi-stakeholder processes. The NTIA launched one such initiative on Mobile Application Transparency in 2012. The result was a draft Code of Conduct for mobile short-form notices. That draft defines a standard short-form privacy notice for apps, which is not to be a substitute for a longer, complete privacy policy.

Multi-stakeholder processes are viewed by some as an improvement over industry self-regulation, in that more stakeholders have a voice. The development of a privacy Code of Conduct through a multi-stakeholder process is thought to facilitate the involvement of media, citizens, and academics, as well as lobbyists, non-profits, and industry. This was the first multi-stakeholder process conducted by the NTIA, and was considered a learning process. The NTIA MSHP included meetings every few weeks in Washington, DC that were open to the public and allowed for remote participation by calling in or viewing a webcast. Participants included lobbyists from companies involved in app development, representatives of consumer-advocate non-profits, and privacy lawyers representing interested companies.

The NTIA multi-stakeholder group struggled with the role of usability testing in drafting the policy. While a usability subgroup was initiated and met several times, no consensus was reached on what should be tested, or by whom. Some stakeholders argued that it had been so difficult to reach consensus on the wording of the code that they were unwilling to submit it to user testing. User testing risked dragging out the process longer than needed. Finally, the subgroup did not perform any user or usability studies. Some participants argued that usability was never a goal of the process.

The user study reported in this paper was initiated and run independently of the usability group, by our own research group at a university. As a participant in the user-study subgroup, we became aware of the practical issues in initiating a user study, and realized that if a user study were to be done, it would need to be done independently of the group, with our own design, initiative, and funds. Our goal was to examine one portion of the notice, in particular the understandability of the wording suggested for the short form notices. If our study found that there were problems with understandability, we hoped that this would influence the Code and the process, and allow the selection of improved terminology.

A. NTIA MSHP draft wording

The NTIA MSHP draft includes seven categories of information to include in in-app privacy disclosures. It also includes eight categories of entities with which data might be shared. The draft includes short definitions for all information types and entities – referred to throughout the paper as the "parenthetical" text – shown in parentheses below. We tested the wording used in the NTIA MSHP draft code published on April 29, 2013.[1] We deliberately did not change, add, or in any way modify the wording or punctuation.

The categories for data types are:

- Biometrics (information about your body, including fingerprints, facial recognition, signatures and/or voice print.)
- Browser History and Phone or Text Log (A list of websites visited, or the calls or texts made or received.)
- Contacts (including list of contacts, social networking connections or their phone numbers, postal, email and text addresses.)
- Financial Information (Includes credit, bank and consumer-specific financial information such as transaction data.)

[1] http://www.ntia.doc.gov/files/ntia/publications/mobileappdraftapril29_2013_draft1b_fs.pdf

- Health, Medical or Therapy Information (including health claims and information used to measure health or wellness.)
- Location (precise past or current location and history of where a user has gone.)
- User Files (files stored on the device that contain your content, such as calendar, photos, text, or video.)

The categories for entities with which data was shared are:

- Ad Networks (Companies that display ads to you through apps.)
- Carriers (Companies that provide mobile connections.)
- Consumer Data Resellers (Companies that sell consumer information to other companies for multiple purposes including offering products and services that may interest you.)
- Data Analytics Providers (Companies that collect and analyze your data.)
- Government Entities (Any sharing with the government except where required or expressly permitted by law.)
- Operating Systems and Platforms (Software companies that power your device, app stores, and companies that provide common tools and information for apps about app consumers.)
- Other Apps (Other apps of companies that the consumer may not have a relationship with)
- Social Networks (Companies that connect individuals around common interests and facilitate sharing.)

V. Methodology

We conducted an online survey using Amazon's Mechanical Turk crowdsourcing service (MTurk)[2] over a two-week period in May 2013. Participants were recruited with the text, "Give us your opinion about information about smartphone apps. This should take 15-25 minutes," and paid $1 for completing the survey.

Previous research has demonstrated that offline experimental results can be successfully replicated using MTurk [38]. Furthermore, while MTurk workers are younger and more technically savvy than the general US population, MTurk has been shown to provide a more diverse sample than a university lab survey [39]–[41]. Using MTurk has allowed us to conduct our study with a larger and more diverse sample than would otherwise have been possible.

We also invited NTIA MSHP members to participate in the same study. MSHP members answered two additional questions about their role in the process. MSHP participants were not compensated. The process for participating in the NTIA is open, but requires a time commitment and dedication to attend and participate in the meetings. These participants are considered experts, since they are familiar with objectives of the NTIA and have worked to shape the draft Code. We advertised the study to MSHP members through announcements by email and a brief presentation at one of their meetings. Response was limited, with only 4 experts (out of 25-50 participants) taking the survey. While we present their responses, we make no statistical claims about the results.

[2]https://www.mturk.com/

A. Survey Design

Our survey presented participants with a sequence of ten randomly-ordered smartphone-app scenarios. In each scenario, we described the app's purpose, what data it collects, and with which entities it shares that data. Some scenarios also included an explanation about why the data is collected. We then asked participants to categorize both the data being collected and the entities with which it is shared, according to the NTIA categories. An example scenario is below. All ten scenarios are provided in the appendix to the Technical Report, which is available online at https://www.cylab.cmu.edu/files/pdfs/tech_reports/CMUCyLab13011.pdf.

> The Fitness app integrates with your FitMonitor (FitMonitor is a special pedometer and activity monitor, purchased separately) to allow you to track and improve your fitness activities and level.
>
> Fitness app will collect information on how many steps you have taken, how long you've slept, and allow you to enter you weight and body fat.
>
> Fitness app will notify sports and health companies if you achieve certain goals, and these companies will send you valuable coupons as awards.

We attempted to represent every data category and every entity category from the NTIA draft in our scenarios. Our scenarios were designed to be realistic. Many scenarios were based on real apps or websites, though we changed the names and adjusted the wording in order to avoid confusion if the participant was already familiar with the real app. In some cases, we took descriptions of apps from the app stores or web sites. We guessed with whom data would be shared, as the apps typically did not reveal this. Our scenarios were more concise, explicit, and specific than typical privacy policies. In three cases, we used the names of real companies — Apple, Facebook, and Google — in order to investigate whether participants considered them to be social networks or operating systems. We included several scenarios that may be considered privacy sensitive. Two scenarios described collecting financial information and another described collecting the user's weight. The "FindMyKid" app allowed a user to set up tracking on someone's phone without that person being aware; such an app could be used by stalkers or abusive partners with physical access to a victim's phone.

B. Data and Entity Categories

After participants read the scenario, they were asked to categorize each type of data and third-party entity with which the data would be shared, based on the NTIA MSHP short-form terms. We presented the categories using the exact same wording, in the same order, as used in the NTIA MSHP draft,

published April 29, 2013. We also added "None of the Above" and "Not Sure" options.

The NTIA provides both names and explanatory text for each category. In order to gain a better understanding of the utility of including this explanatory text, we conducted our study as a between-subjects survey. Participants in the *terms only* condition were shown only the category names in each scenario; participants in the *parentheticals* condition were also shown the NTIA's explanatory text for each category.

We designed our online survey after conducting eight in-person pilot tests, in which the survey-taker walked through the survey with the researcher and thought out loud. These pilots allowed us to refine our study design. For example, in these pilot surveys, we found that participants were skeptical about the scenarios giving them complete information about what data would be shared. As a result, they were apt to make inferences about additional types of data that might be shared. Therefore, we designed the survey so that participants would select a data or entity option only for elements mentioned explicitly in the scenario. Furthermore, we added a notice on every page stating, "The scenarios describe the data collection and sharing completely, so you do not need to guess anything outside of what is described." We also included two open-ended questions that were used as an attention check for quality results.

C. Data Analysis

Each of our participants was shown a sequence of ten scenarios; each scenario had at least one data item and at least one third-party entity with which data is shared. Participants were asked to classify each data item and each entity according to the NTIA categories, or as "None of the Above" or "Not Sure." In all, participants were asked to make 52 categorizations. The data type items we asked participants to categorize are shown in the second column of Table I, and the third-party entities are shown in Table II.

We cannot determine how many of our participants were "correct" in each scenario, because there is no ground-truth on which to base that assessment. This is the result of the stake-holder process, in which there were concessions but not necessarily agreement on the terms and their meanings. Thus, there is no way to determine whether a given response is inherently correct or incorrect. Given this lack of general correctness, instead our analysis focuses on how consistently our participants categorized the data items and entities. For each data item and entity, we considered the most-commonly selected category to be the *winner*. We then looked at the percentage of participants who selected the *winning* category for each data item and entity, and we call this percentage the *common understanding* for that data item or entity.

We classify each data item and each entity as being either *low common understanding* or *high common understanding*. A data item or entity in which more than 60% of our participants agreed on its categorization is considered to be *high common understanding* (that is, more than 60% of participants categorized it as its *winning* categorization). A data item or entity with 60% or lower categorization agreement is considered to be *low common understanding*.

VI. Study Results

Our study found that participants and the NTIA experts had a low common understanding of many of the terms used in the NTIA MSHP notice. We begin with a description of our participants and then summarize our main findings. Detailed results of the study can be found in Tables I and II. Breakdowns of how participants voted for each element can be found in the appendix to the Technical Report, which is available online at https://www.cylab.cmu.edu/files/pdfs/tech_reports/CMUCyLab13011.pdf.

A. Participants

The four NTIA MSHP participants in our study, whom we call our *expert participants*, were a diverse group. They each held different professions and represented different stakeholders in the NTIA process; we do not report their demographics to preserve their anonymity. Expert participants were evenly split between our two conditions; because we had only two expert participants in both conditions, we do not report differences based on these conditions for expert participants.

For our MTurk participants, we analyzed data only for participants in the United States who had completed the survey, and we excluded participants who entered gibberish answers for open-text fields that were used as an attention check. This left us with 791 MTurk participants (375 parenthetical and 416 term-only). The data was collected in two batches, one of 503 responses and one of 288 responses. The second batch included three data entities accidentally omitted from the first. The data entities were: Sports and Health Companies in the Fitness scenario and AdMeMetric in the Salsa scenario; these are indicated in Table II. We combine the results from these two batches, except when discussing the three questions that had only 288 responses.

51% of the MTurk respondents were female. Participants ranged in age from 18 to 73 years, with a mean of 33 and a standard deviation of 11 years. Participants took an average of 17 (median 15) minutes to complete the survey. Every US State was represented. Participants were generally educated: 38% have a Bachelors degree, and another 30% have some college. 82% own a smartphone.

B. User Study Findings

The NTIA MSHP has selected several categories of data sharing about which mobile users should be informed on short-form privacy notices. Our investigation looked at user and expert understanding of these categories. Our survey found that the categories were not well understood by our participants. Of the 52 examples of data sharing given in our scenarios, participants showed low (less than 60%) common agreement for 23 of them. Furthermore, our expert participants also disagreed among themselves on how to categorize some of the examples, and had different majority responses from the study participants for 13 examples. We find that the *Biometrics* and *Health, Medical or Therapy Information* categories were especially prone to disagreement. Further, participants struggled to categorize many of the third-party entities.

Scenario	Data	Expert Response	Winning Participant Response	Parenthetical[1]	Term only[1]	p-value
HipClothes	Inseam	Biometrics (2)	Biometrics	69.1	45.9	<.001*
	Waist Size	Biometrics (2)	Biometrics	69.6	46.4	<.001*
	Clothing Preference	None (3)	None	48	38	<.001*
	Location	Location (4)	Location	91.7	89.9	.494
Salsa	Call History	Browser History (4)	Browser History	88.5	87.5	.463
	Text History	Browser History (4)	Browser History	89.3	90.1	.184
	Video History	Browser History (4)	Browser History	51.5	70	<.001*
	Games Played	Browser History (3)	Browser History	45.9	50.5	.021*
	Photos	User files (3)	User Files	77.6	69.2	.005*
SuperTax	Photo of W2	Financial Information (3)	User Files	59.2	75.5	.001*
	Salary	Financial Information (4)	Financial Information	92.3	93.3	.502
	Interest Income	Financial Information (4)	Financial Information	92.5	91.8	.066
Fitness	Steps Taken	Health (2)	Biometrics	40.3	46.2	.225
	How Long Slept	*Health (4)*	Biometrics	39.7	44.2	.148
	Weight	*Health (4)*	Biometrics/Health	54.1	50.2	<.001*
	Body Fat	*Health (4)*	Biometrics/Health	53.3	49.5	.005*
EasyApply	Work History	*None (3)*	None /Financial Information	33.3	34.4	<.001*
	Medical Insurance	Health (3)	Health	85.9	81	.161
	Medical Payments	Health (4)	Health	59.7	52.2	.127
	Number of Children	None (3)	None	41.1	35.1	<.001*
	Marital Status	None (3)	None	43.5	35.1	<.001*
	Income	Financial Information (4)	Financial Information	88.5	91.6	.063
CallCalendar	Call Time	Browser History (4)	Browser History	91.2	86.8	.222
	Call Duration	Browser History (4)	Browser History	90.1	86.3	.189
	Name from Contact List	Contacts (3)	Contacts	71.2	82.5	<.001*
GoodDriver	GPS Location	Location (4)	Location	94.1	94.7	.788
	Gyroscope Bumps	None (3)	None	33.6	33.9	.252
FindMyKid	Location	Location (4)	Location	94.1	94.7	.176
iTunes	Credit Car Info	Financial Information(3)	Financial Information	96	92.3	.304
	Song and Artist Names	*None (3)*	User Files	57.1	53.1	.443
Bookstore	Book Title	None (4)	None	34.4	36.1	.502
	Home Address	None (2)	Location	49.1	58.7	.008*
	Credit Card	Financial Information(4)	Financial Information	94.1	91.1	.092

[1] Participant level of common understanding for winning term by condition (% who selected the winning participant response).
* Difference between conditions is significant at p<.05 with χ^2 test Benjamini and Hochberg FDR correction.

TABLE I. DATA TYPE CATEGORIES SELECTED FOR EACH TERM BY NTIA EXPERTS AND MTURK PARTICIPANTS. FOR THIS TABLE, THE CATEGORIES "HEALTH, MEDICAL OR THERAPY INFORMATION" HAS BEEN ABBREVIATED TO "HEALTH" AND "BROWSER HISTORY AND PHONE OR TEXT LOG" TO "BROWSER HISTORY." IN THE EXPERT COLUMN, WE SHOW ALL CATEGORIES SELECTED BY TWO OR MORE EXPERTS, WITH THE NUMBER OF EXPERTS THAT SELECTED EACH CATEGORY IN PARENTHESES. THE TERMS IN WHICH THE MAJORITY OF EXPERTS AND PARTICIPANTS DIFFERED ARE IN ITALICS. IF THE CONDITIONS IN THE PARTICIPANT STUDY HAD DIFFERENT WINNERS, BOTH ARE SHOWN IN THE PARTICIPANT COLUMN.

The main finding of this study is that the current set of NTIA categories does not appear to offer a high level of transparency for users. The lack of common understanding, even among experts, also suggests that app developers may have trouble generating accurate notices using these terms and definitions. Next, we will discuss our main findings and our recommendations.

Parentheticals Help (Sometimes). In most cases, the difference between the parenthetical condition and the term-only condition was not significant. When it was significant, the parenthetical usually resulted in greater agreement with the most-popular category. However, this was not always the case; some parentheticals appeared to confuse our participants. For example, the parenthetical text for *Browser History and Phone or Text Log*, *User File*, and *Location* appear to need some improvement to make them more useful to users.

Better Definitions Are Needed. Some categories were not well understood, either by participants or by NTIA experts. Therefore, we recommend that the Code provide further guidance on how to interpret the categories. This may include definitions and examples, including edge cases. In particular, guidance is needed for all of the third-party entities except *Government Entities*, as well as the categories *Biometrics* and *Health, Medical or Therapy Information*. Further, experts should clarify whether location includes only information from sensors (such as GPS) or user-entered information (such as home address).

Ambiguous Data Items Need Clarification. Several types of data items were confusing to participants. Some data items could reasonably be classified in two categories (e.g., a photo of a W-2 is both a user file and financial information). This typically resulted in low common understanding. Furthermore, the Code of Conduct does not specify whether both categories should be listed or how one should be chosen. Some data items require an understanding of the platform architecture in order to classify them correctly (e.g., whether a contact name is stored in a call log or in a user file). As a result, app

Scenario	Data	Expert Response	Winning Participant Response	Parenthetical[1]	Term only[1]	p-value
HipClothes	OtherClothingStores	*None (3)*	Consumer Data Reseller /None	31.5	33.3	<.001*
Salsa	Advertising Companies	Ad Networks (4)	Ad Networks	80.5	79.2	.520
	AdmeMetric[2]	*Consumer Data Reseller (3)*	Consumer Data Reseller	43.8	38	.086
SuperTax	State Agency	Government Entity (4)	Government Entity	93.9	96.2	.465
	Federal Agency	Government Entity (4)	Government Entity	94.7	95.4	.518
Fitness	Sports Companies[2]	*None (3)*	Consumer Data Reseller	38.4	26.8	.027
	Health Companies[2]	*None (3)*	Consumer Data Reseller	31.5	24.6	.022*
EasyApply	State Agency	Government Entity (4)	Government Entity	92	93.3	.208
CallCalendar	Carrier	Carrier (4)	Carrier	90	88.2	.173
	Google Calendar	Other Apps (3)	Other Apps	47.1	51	.066
GoodDriver	Traffic Data Company	None (2)	Data Analytics	59.7	58.4	.770
	Car Insurance	*None (4)*	Consumer Data Reseller	35.7	26	<.001*
	Car Rental	*None (4)*	Consumer Data Reseller	36.3	25.7	<.001*
FindMyKid	Parents Phone	None (3)	None	34.4	46.6	.034
	Local Police	Government Entity (4)	Government Entity	80	85.3	.333
iTunes	Facebook	Social Network (3)	Social Network	89.6	92.1	.714
	Apple iCloud	OS and Platforms (2), None (2)	OS and Platforms	37.9	34.9	.799
Bookstore	Facebook	Social Network (3)	Social Networks	88.8	90.6	.566
	GreatReading	Social Network (2), Other Apps (2)	Other Apps	37.6	40.1	.410

[1] Participant level of common understanding for winning term by condition (% who selected the winning participant response).
[2] 288 Responses Only
* Difference between conditions is significant at p<.05 with χ^2 test and Benjamini and Hochberg FDR correction.

TABLE II. THIRD-PARTY ENTITIES CATEGORIES SELECTED FOR EACH TERM BY NTIA EXPERTS AND MTURK PARTICIPANTS. IN THE EXPERT COLUMN, WE SHOW ALL CATEGORIES SELECTED BY TWO OR MORE EXPERTS, WITH THE NUMBER OF EXPERTS THAT SELECTED EACH CATEGORY IN PARENTHESES. THE TERMS IN WHICH THE MAJORITY OF EXPERTS AND PARTICIPANTS DIFFERED ARE IN ITALICS. IF THE CONDITIONS IN THE PARTICIPANT STUDY HAD DIFFERENT WINNERS, BOTH ARE SHOWN IN THE PARTICIPANT COLUMN.

developers may correctly categorize a data type, but users may not understand the categorization.

In several cases, participants who saw the parenthetical text had less agreement than those who saw only the terms, indicating that the short phrases created confusion instead of clarification. In the case of home address, participants who saw the parenthetical were less likely to select *Location*, and were more likely to say *Not Sure* or *None*. In the case of video history, users who saw the parenthetical text may have been attracted to the word "video" in the *User File* description, and therefore choose that category over *Browser History and Phone or Text Log*.

For improved transparency on ambiguous or poorly understood data types, we recommend that implementors of the short-form specify the data being collected. For example, a short form notice with the text "Health, Medical or Therapy Info: how many steps you have taken, how long you've slept, weight, and body fat" may be more clear to users than "Health, Medical or Therapy Info." The specificity would alleviate the problems described above with ambiguous data types. Future research should investigate whether specific information is better understood, and whether implementors of a short-form notice should specifically say what is being collected instead of, or in addition to, the parenthetical text.

Third-Party Entities Are Poorly Understood. Many of the third-party entity categories were confusing to participants. Our results show that participants struggled with many of the third-party entities, except *Government* and *Carriers*. In particular, participants categorized six entities as *Consumer Data Resellers* while the experts only categorized one as such, typically choosing *None* instead. It may be that participants used this as a fallback choice for entities they didn't understand, while the experts had a much narrower definition in mind.

On the other hand, specificity about third-party entities will only be helpful if users recognize the name of the entity. Previous research suggests that users are not familiar with the names of advertisers, data resellers, or analytics companies [2], [35]. Further research is needed on describing third-party entities in a transparent way.

Uncategorized Data and Entities. There are some privacy-sensitive data that do not fit into any of the existing categories (and therefore need not be indicated in a short-form notice). These include identifying information such as user name, phone id, or SSN. Since not all data sharing falls into a category covered by the short-form notice requirements, the app may be sharing data without notifying the user through the short form. Our results show that participants did not often categorize data and entities as *None*, and preferred to place data in one of the categories. This suggests participants believe the categories encompass all possibilities. Therefore, information about the smartphone notices should emphasize that the short form does not notify users about all types of data sharing.

Further User Testing is Needed. Our study is a concrete first step which indicates that more work is needed to develop a well-understood notice with categories and definitions that will be generally understood by American smartphone users. By providing realistic scenarios and asking survey participants

to categorize data items and entities with which data is shared, our work highlights that the categories are not well understood. However, this is not a typical task flow for users, and we did not test actual short-form notices. However, if the NTIA MSHP had adopted a similar approach of using case studies to understand and categorize data sharing, it is likely they would have developed more understandable terms and definitions. In fact, when similar examples were raised in meetings the group moved on without reaching a consensus.

VII. Limitations

This survey is designed to measure whether participants understand the NTIA categories by giving them an explanation of an app, and an explanation of the data shared, including such details as with whom the data is shared and the purpose of sharing the data. Participants may see more information than they would in practice. Our results for understanding, therefore, may be an overestimate of true understanding in practice. Further, as stated above, while we can measure the extent to which participants agree on how to categorize a given data item or entity, it is impossible to determine whether that categorization is "correct."

The task presented to survey participants more closely resembles a realistic task for an app developer than a user. A more realistic user task might be to provide a notice that uses the terms from the Code and to ask users what data they think an app is collecting and with what entities they believe it is shared. However, this is actually an even harder task because each data category could potentially cover many types of data, and it is not necessarily possible to infer what data is collected from a very brief description of an app.

This survey is limited to testing the particular terminology defined by the NTIA code. While the results indicate some categories are poorly understood, we do not test alternate wordings. Therefore, we are unable to offer better terminology; that may be an area for future work.

Furthermore, while we tried to present a broad swath of scenarios, we could not create a study that would present all possible scenarios to participants. There may be many more types of data that are ambiguous to users, or examples that are more clear than those in this survey.

Our pay rate of $1 for a 17-minute survey was well above the mturk rates studied by Buhrmeister et. al, which showed that the lower rates did not effect the quality of results [42]. However, it is possible that the low pay could have impacted the quality of results.

Although we did extensive in-person piloting, we were not able to pilot extensively the survey with Amazon mturk participants. This was due to our deadline of completing the work with enough time to inform the NTIA MSHP before the final meeting. As mentioned in the results section, the data was collected in two batches, in which the second batch included three entities that were not in the first batch. Due to time constraints, we did not discard the first set. A χ^2 test between the two batches found no significant differences for the other questions. Therefore, we have reason to believe that combining the two batches did not impact the results of the questions.

Our recommendations for usable privacy and security practitioners are based on one case study. Although they are drawn from several informal discussions with other participants (such as personal conversations over the phone), we do not present them as results from a qualitative study.

VIII. Discussion

We released a technical report of our work on July 17, 2013, one week before the final NTIA MSHP meeting on July 25th. This technical report showed that the terminology in the short-form notice was not well understood, and further research was needed. However, by this point, participants were ready to reach consensus on the Code of Conduct and conclude the project. Although our study was discussed by the group, by then it was too late in the process to influence the Code very deeply. That said, we believe our study did have an impact, as future discussions indicated that user studies were planned [43]. Unfortunately, the process has already concluded, and the Code of Conduct has been announced, without plans to reconvene or address the usability issues. We fear that despite the best intentions of the participants, this will lead either to adoption of a short-form notice that does not meet its goals due to usability issues, or to app developers finding the notice flawed and therefore not adopting the voluntary Code.

Here we provide lessons learned, particularly aimed at academics or experts on usability who believe that regulation around technology should consider users and the human element. These are based on our own experience and personal off-the-record discussions with several stakeholders. We distill our lessons for academics and usability experts who also wish to avoid policies with requirements that are known to be unusable. This is organized into two subsections. The first subsection describes specific issues that hindered the integration of usability studies into the NTIA MSHP. We describe these issues with the goal of illuminating some gaps between academic HCI experts and the policymakers. We then offer some concrete suggestions for usability experts who wish to participate in multi-stakeholder public-policy-making.

A. Issues that hindered usability testing

In this section we describe specific issues that occurred during the NTIA process that led to adopting a Code of Conduct that was not well understood by the users in our study.

Disagreement about what 'usabilty' is. The main issue with conducting usability studies was that the stakeholders did not agree on what 'usability' meant. Although the stakeholders often recognized the need for user studies, they had different opinions about what should be studied, how it should done, and what the results would mean. This is largely a result of the multi-stakeholder process, in which different stakeholders had different objectives and priorities, based on their experiences and whom they were representing.

For example, some stakeholders representing app developers felt that if usability tests showed some users were concerned by the notice and therefore did not download an

app, this indicated a failure in usability. In contrast, some consumer privacy advocates argued that the notices should lead consumers to refuse the data collection practices of apps and download fewer apps. This difference in opinion may be familiar to those who have worked on notifications in other areas, such as authentication or P3P.

Other debates about usability included the role of icons in the notice, and whether icons could stand alone, or with text, or whether icons should be allowed at all [44]. Another debate was about which entities and elements the app needed to show on the notice: either only those the app collects or shares with, or the entire list with an indication that some things are not shared. Some stakeholders felt that usability tests should address these questions, while others felt the questions were not relevant or were not the primary issues of interest. Our study did not address these issues, but we agree they should be examined.

Cost of usability studies. Part of the delay in starting usability studies was that it was difficult to resolve who should pay for the studies, and who should carry them out. The NTIA process itself did not have a budget for usability studies, so in order to pay usability consultants or private usability firms, some stakeholders would need to volunteer the funds. Although a stakeholder volunteered to search for usability consultants and request prices, it was difficult to get an estimate without knowing what would be tested. Furthermore, it was not clearly defined whether financial contributions from stakeholders would give them more control over the tests. The final cost of our study was under one thousand dollars, not including the value of graduate students' time to implement the study, and was paid by our lab's funds.

Process fatigue. After a year, many of the stakeholders were eager to complete the project. Although fatigue with the process may have contributed positively to the ability to come to a consensus, it also meant that the stakeholders were not willing to wait for usability results. This may be a different perspective than that of academics, who are often willing to dive into an interesting problem for several years.

Everyone is expected to have a bias. It has been said that policy makers at the federal level expect everyone to have a position. For example, several participants in the NTIA MSHP felt that future processes should request that all participants submit position papers. Academics may feel they don't take a 'position'; they strive to be neutral and let the results of the research stand as facts that support their claims. However, in controversial areas such as privacy, academics should be prepared to describe their position. In our case, our position was that the Code of Conduct should be usable; that both smartphone device users and app developers should understand the notice and the terms used.

B. Recommendations

In this experience, we found that although the drafters of the Code of Conduct generally recognized the value of user studies, they were unable to implement those studies. Our independent user study confirmed that categorizations described in the Code were confusing and suggested that further user studies could help create a more understandable privacy notification.

Engage early. Our independent study confirmed that usability tests were needed. We recommend that other researchers who have the resources can and should conduct user-tests to inform public policy, as it may not happen otherwise. We do not recommend waiting until stakeholders come to agreement about what to test. With the benefit of hindsight, we should have run our tests sooner to inform the process at an earlier stage.

Furthermore, we released a technical report before the final NTIA MSHP meeting, so that the results of the study would be available for the participants. We did not wait for publication in an academic journal or conference, as this may have delayed the results beyond the point of impact.

Our technical report could have been more useful if we had included a one-page executive summary. This may have been more relevant and useful to stakeholders and journalists than a full-length paper for understanding the issues within their time constraints.

Impact versus incentives. We recognize that academics may have little incentive to put their resources toward such studies, which may have more value to policymakers or a working group than to academics or reviewers. Indeed, our attempts to publish this paper in an academic conference were initially thwarted by reviewers who felt the results of the study did not make a significant contribution to the field. Ultimately, we refocused the paper as a case study before submitting it to this workshop. However, we believe that engaging with public policy can help prevent requirements with poor usability from being written into regulation. This may increase the impact of our research, as a community, in the long run.

ACKNOWLEDGEMENTS

This research was funded in part by NSF grants DGE0903659, CNS1116934, and CNS1012763 and John and Claire Bertucci Fellowship.

REFERENCES

[1] L. Cranor, *Web privacy with P3P*. O'Reilly Media, Inc., 2002.

[2] P. G. Leon, B. Ur, R. Balebako, L. F. Cranor, R. Shay, and Y. Wang, "Why johnny can't opt out: A usability evaluation of tools to limit online behavioral advertising," in *Proc. CHI*, 2012, pp. 589–598.

[3] A. McDonald and J. Peha, "Track gap: Policy implications of user expectations for the 'Do Not Track' internet privacy feature," in *TPRC*, 2011.

[4] P. G. Kelley, L. Cesca, J. Bresee, and L. F. Cranor, "Standardizing privacy notices: an online study of the nutrition label approach," in *Proc. of CHI 2010*, 2010. [Online]. Available: http://doi.acm.org/10.1145/1753326.1753561

[5] *Evolution of a Prototype Financial Privacy Notice: A Report on the Form Development Project*. FTC and Kleimann Communication Group, Inc, 2006. [Online]. Available: http://www.ftc.gov/privacy/privacyinitiatives/ftcfinalreport060228.pdf

[6] A. Levy and M. Hastak, "Consumer comprehension of financial privacy notices: A report on the results of the quantitative testing," *Federal Trade Commission*, pp. 62 890–62 994, Dec. 2008.

[7] "Final model privacy form under the Gramm-Leach-Bliley act; final rule," *Federal Register*, vol. 74, no. 229, pp. 62 890–62 994, Dec. 2009.

[8] M. Hastak and M. Culnan, "Online behavioral advertising "icon" study," *Future Of Privacy Forum*, 2010. [Online]. Available: http://futureofprivacy.org/final_report.pdf

[9] L. Norden, W. Quesenbery, and D. C. Kimball, "Better design, better elections," Brennan Center for Justice at New York University School of Law http://www.brennancenter.org/publication/better-design-better-elections, 2012.

[10] C. Lewis and J. Treviranus, "Public policy and the global public inclusive infrastructure project," *interactions*, vol. 20, no. 5, pp. 62–66, 2013.

[11] A. Cavoukian, "Privacy and biometrics," Information and Privacy Commissioner, Ontario, September 1999.

[12] S. Prabhakar, S. Pankanti, and A. Jain, "Biometric recognition: security and privacy concerns," *IEEE Security & Privacy*, vol. 1, no. 2, pp. 33–42, 2003.

[13] A. M. McDonald and L. F. Cranor, "Americans' attitudes about internet behavioral advertising practices," in *Proc. WPES*, 2010.

[14] B. Ur, P. G. Leon, L. F. Cranor, R. Shay, and Y. Wang, "Smart, useful, scary, creepy: perceptions of online behavioral advertising," in *Proc. SOUPS*, 2012.

[15] P. G. Leon, B. Ur, Y. Wang, M. Sleeper, R. Balebako, R. Shay, L. Bauer, M. Christodorescu, and L. F. Cranor, "What matters to users? factors that affect users' willingness to share information with online advertisers," in *Proc. SOUPS*, 2013.

[16] D. J. Solove, "Five myths about privacy," Washington Post, June 13, 2013.

[17] C. Arthur, "Is your private phone number on facebook? probably. and so are your friends'," The Guardian, October 6, 2010. [Online]. Available: \url{http://www.guardian.co.uk/technology/blog/2010/oct/06/facebook-privacy-phone-numbers-upload}

[18] D. Smilkov, D. Jagdish, and C. Hidalgo, "Immersion," https://immersion.media.mit.edu/, 2013.

[19] B. Isaacson, "Immersion, An MIT Media Lab Creation, Uses Email Metadata To Map Your Connections," Huffington Post, http://www.huffingtonpost.com/2013/07/10/immersion-email-metadata_n_3567984.html, July 10, 2013.

[20] D. Solove, "'I've got nothing to hide'and other misunderstandings of privacy," *San Diego law review*, vol. 44, 2007.

[21] M. Marlinspike, "Why 'I have nothing to hide' is the wrong way to think about surveillance," Wired, June 13, 2013.

[22] J. Valentino-Devries, J. Singer-Vine, and A. Soltani, "Websites vary prices, deals based on users' information," Wall Street Journal, December 24, 2012.

[23] G. J. Annas, "HIPAA regulations, a new era of medical-record privacy?" *New England Journal of Medicine*, vol. 348, no. 15, April 2003.

[24] E. Toch, J. Cranshaw, P. H. Drielsma, J. Y. Tsai, P. G. Kelley, J. Springfield, L. Cranor, J. Hong, and N. Sadeh, "Empirical models of privacy in location sharing," in *Proc. Ubicomp*, 2010, pp. 129–138.

[25] M. Benisch, P. G. Kelley, N. Sadeh, and L. F. Cranor, "Capturing location-privacy preferences: quantifying accuracy and user-burden tradeoffs," *Personal Ubiquitous Comput.*, vol. 15, no. 7, pp. 679–694, Oct. 2011.

[26] I. Muslukhov, Y. Boshmaf, C. Kuo, J. Lester, and K. Beznosov, "Understanding Users' Requirements for Data Protection in Smartphones," *ICDE 2012*, pp. 228–235, Apr. 2012. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6313685

[27] A. Felt, S. Egelman, and D. Wagner, "I've got 99 problems, but vibration ain't one: A survey of smartphone users' concerns," in *Proc. SPSM*, 2012.

[28] "Fact Sheet 39: Mobile Health and Fitness Apps:What Are the Privacy Risks?" *Privacy Rights Clearinghouse*, 2013. [Online]. Available: https://www.privacyrights.org/fs/fs39/mobile-apps

[29] "Public split over impact of NSA leak, but most want Snowden prosecuted," A Pew Research Center/USA TODAY Survey, June 17, 2013.

[30] B. Krishnamurthy and C. E. Wills, "Characterizing privacy in online social networks," in *Proc. of WOSN '08*, 2008. [Online]. Available: http://doi.acm.org/10.1145/1397735.1397744

[31] R. Gross and A. Acquisti, "Information revelation and privacy in online social networks," in *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*. ACM, 2005, pp. 71–80.

[32] J. Urban, C. Hoofnagle, and S. Li, "Mobile phones and privacy," *UC Berkeley Public Law Research Paper*, 2012.

[33] A. Felt, E. Ha, S. Egelman, A. Haney, E. Chin, and D. Wagner, "Android permissions: User attention, comprehension, and behavior," *Proc. of SOUPS*, 2012.

[34] P. Kelley, L. F. Cranor, and N. Sadeh, "Privacy as part of the app decision-making process," in *Proc. of CHI 2013*, 2013.

[35] R. Balebako, J. Jung, W. Lu, L. Cranor, and C. Nguyen, "Measuring user confidence in smartphone security and privacy," in *Proc. SOUPS*, 2013.

[36] A. Felt, S. Egelman, M. Finifter, D. Akhawe, and D. Wagner, "How to ask for permission," *HOTSEC 2012*, 2012.

[37] *Consumer Data Privacy In A Networked World: A Framework For Protecting Privacy And Promoting Innovation In The Global Digital Economy*. White House, 2012. [Online]. Available: http://www.whitehouse.gov/sites/default/files/privacy-final.pdf

[38] G. Paolacci, J. Chandler, and P. Ipeirotis, "Running experiments on amazon mechanical turk," *Judgment and Decision Making*, vol. 5, no. 5, pp. 411–419, 2010.

[39] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data?" *Persp. Psych. Sci.*, vol. 6, no. 1, pp. 3–5, 2011.

[40] P. G. Ipeirotis, "Demographics of Mechanical Turk," New York University, Tech. Rep. CeDER-10-01, 2010.

[41] A. J. Berinsky, G. A. Huber, and G. S. Lenz, "Using Mechanical Turk as a subject recruitment tool for experimental research," 2011.

[42] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical turk a new source of inexpensive, yet high-quality, data?" *Perspectives on Psychological Science*, vol. 6, no. 1, pp. 3–5, 2011.

[43] S. Nayak, "What's next for the NTIA mobile app transparency code?" TrustE blog, http://www.truste.com/blog/2013/08/01/ntia-mobile-app-transparency-code, Aug. 1 2013.

[44] "NTIA app group inching toward usability testing," Politico.com, http://www.politico.com/morningtech/0113/morningtech9850.html, Jan. 18 2012.