Putting Disclosures to the Test

an FTC Workshop

STAFF SUMMARY

November 2016

Table of Contents

1. Introduction

On September 15, 2016, the Federal Trade Commission convened a public workshop, Putting Disclosures to the Test, that examined ways of testing and evaluating the effectiveness of disclosures in communicating a wide range of information that consumers need to make informed decisions in the marketplace. Disclosures may be delivered offline or online through icons, product labels, short text, long text, audio or video messages, interactive tools, and other media. The FTC focuses on disclosures that affect consumer welfare such as disclosures that inform consumers about the risks from using certain products, or disclosures necessary to limit or qualify marketing statements in order to prevent deception. Disclosures may inform consumers about the choices they have and allow them to make

“

“

They compared performance between participants who were shown the ads with no labels, and labels with varying wording, size and color, and position on the page. According to Dr. Hyman, because small changes to ad labels can make a big difference in ad recognition, it is important to conduct experiments to test what is most effective.

Finally, Rebecca Balebako[25] presented two experiments[26] one in which participants downloaded and used a quiz app on their own smartphones, and one in which participants virtually downloaded and used the same app on their computer screens. Participants in both experiments were divided randomly into groups in which they were presented with a privacy disclosure under different conditions. The disclosure was presented before downloading the app or at various times while using the app. After completing the quiz, participants answered memory questions about the app, including questions about the privacy notice. Dr. Balebako said that in both experiments, participants presented with the notice prior to downloading the app were significantly less likely to answer questions about it correctly than those shown the notice after the app was installed. She emphasized that timing made a difference in disclosure effectiveness and that the online and field study results were consistent.

lly

Elizabeth Howlett[28] discussed research[29] on how consumers comprehend front-of-package nutrition labeling. She conducted an online study in which participants were shown food product packages with and without objective and evaluative nutrition icons. Dr. Howlett reported that when participants evaluated a single food item in a noncomparative context, the objective icon was most effective at informing participants about how nutritious the food was. She followed up with a study conducted in a lab that contained grocery store shelves stocked with food products. Dr. Howlett found that when she asked participants to compare, rather than consider in isolation, two similar products, the evaluative icon was most helpful in choosing the healthier food. Thus, she concluded that the context in which people process disclosures makes a difference when evaluating comprehension.

Susan Kleimann[30] discussed the iterative design and evaluation approach[31] she used to develop mortgage disclosures intended to help consumers comprehend information about loans they are considering, compare loans, and choose the best loan for their situation. The disclosures were refined over 18 rounds of qualitative interviews with consumers in English and Spanish. Researchers used a framework known as "Blooms taxonomy"[32] to identify the participants' stages of understanding – knowledge, comprehension, application, analysis, synthesis, and evaluation. Finally, the designs were evaluated in an 858-participant study conducted at 20 locations in which participants were asked to complete comprehension, comparison, and choice tasks using either the existing or proposed (l)-2(e)1DC BT 0 T(e)1or. end ie (i)-2(h(c)4(e)4(-1(e)zTc 0.or)3:-12(c)2(n 4(Ca)-6(e)4(r)3(s)-1(c)-6(opr)3(

people. In addition, he said that previous studies tended to ask every participant a large number of questions due to the cost of recruiting each participant, and that sometimes resulted in nonsensical answers by the end of a long survey. He pointed out that with the availability of crowdsourcing platforms, such as Amazon Mechanical Turk, where one can hire individuals to complete tasks, large numbers of people can be surveyed at a much lower cost and that it is feasible to ask each participant only a small number of questions. Dr. Howlett added that using a crowdsourcing platform, she now completes data collection for a study in a single afternoon for a much lower cost than using a marketing research service. Drs. Goldstein and Howlett said journal reviewers have questioned the quality of Mechanical Turk experiments, but there have been studies that have demonstrated that data quality can be just as good and sometimes better when experiments are carried out on Mechanical Turk rather than with traditional methods.

Dr. Goldstein also noted that he has used mouse tracking successfully in online studies as an inexpensive proxy for eye tracking: "This is basically to look at where people are positioning the mouse on the screen. It is a good proxy of where people are looking on the screen, because people read with their mouse more than they think." He added that the price of eye trackers was also dropping and that it is now possible to do eye tracking through a webcam on a laptop rather than investing in $30,000 eye tracking hardware.

Finally, panelists discussed their experiences evaluating the comprehension of icons. Dr. Howlett noted that sometimes people read more into a simple icon than they should. For example, she described a "health halo" effect for a healthy food icon—for example, people may incorrectly assume that a low sodium food is also low fat. Dr. Reidenberg added that in designing privacy icons it has been difficult to figure out how much weight to assign to the many factors that go into evaluating a privacy policy in order to calculate a single grade that would be meaningful for users.

Dr. Howlett also noted the importance of reference points, similar to Dr. Goldstein's perspective phrases. She said consumers do not know how many grams of trans fat or sodium are healthy or how much exercise it takes to work off a certain number of calories.

5. Impact on decision-making and behavior

The "Impact on Decision-Making and Behavior"[34] panelists discussed surveys, online experiments, observational studies, and field experiments that evaluate the impact that disclosures have on consumers' decision-making and behavior. Panelists examined data breach notifications, privacy notices, restaurant and physical report cards, and payday loan disclosures.

Lillian Ablon[35] presented a consumer survey on data breach notifications.[36] The survey was conducted using the "American Life Panel," a nationally representative online panel of American adults. Ms. Ablon noted that while this research method relies on self-reported data, it allows data collection from a representative group with a high response rate. Using this method, she and her colleagues collected data on how often people recalled receiving breach notifications, and what impact people reported these notifications had on their behavior. Ms. Ablon reported that 26% of respondents recalled receiving a breach notification in the preceding 12 months, and more than half of those said they

conducting a controlled field study with customers at multiple locations, as well as the reasons the lender agreed to participate in the study.

Panelists discussed the various methodologies that they used and that there were advantages and

7. The future of disclosures?

The last panel[50] presented studies that evaluate new approaches or new applications of existing approaches to disclosure design and presentations that suggest ways to make disclosures more efficient and effective. The studies focused on mobile app permissions, privacy notices, and medical study informed consent forms.

Serge Egelman[51] described research in which 133 participants' Android phones were equipped to collect data about app usage and the user's browsing, screen locking, and preference data. The researchers collected 176 million events in which apps accessed sensitive data. About five or six times per week for one month participants were prompted with notices informing them about the type of data an app on their phones had just accessed and asking whether they would have allowed this if given the choice. The researchers developed a model of what data would be collected in an "ask-on-first-use" situation where people are prompted for permission only the first time an app tries to access data (the status quo) and how that differed from the preferences participants expressed in their responses to study prompts. According to Dr. Egelman, they determined that 20% of the time participants would have denied subsequent requests and therefore that the ask-on-first-use model produces a 20% error. He said that the researchers then used machine learning to create a model based on users' behavioral data collected for the study. This model correctly predicted whether users would have granted access about 96% of the time, thus reducing the error rate to about 4%. Dr. Egelman said these results raise questions about whether it would be better to ask for permissions to access sensitive data less frequently, using the machine learning model to make permission decisions the rest of the time, or to prompt users more frequently and risk habituating people.

Tamar Krishnamurti[52] discussed research on shortening informed consent disclosures for medical studies based on feedback about what participants find most important. Researchers began with a 17-page consent form for a clinical trial of an asthma treatment. They used Mechanical Turk workers who were self-reported asthma patients to read a section of the consent form, select the sentences pertinent to making an informed decision about enrolling in a trial, and rate those sentences on how important they were to their decision-making. They then developed a 5-page consent form based on participants' preferences

policy that removed terms known by either 85% or 70% of participants. Then they tested the effectiveness of the short, medium, and long policies. According to Dr. Schaub, they found that all three policies increased awareness of the company’s practices but the shortest policy performed significantly worse than the other policies. In addition, he said that all participants spent about the same amount of time reading each policy, regardless of length. Dr. Schaub also talked about research to develop personalized privacy assistant[55] software for mobile phones that can learn privacy preferences over time.

Panelists noted that some of the approaches to making disclosures more efficient could raise legal and ethical issues. They discussed the question of how to determine the cutoff for what information to present to a consumer and what happens if automated systems make incorrect decisions. Panelists said short or infrequent disclosures could not completely replace full disclosures, but should be viewed as a top layer that would not prevent consumers from going deeper and looking at a full disclosure. They also agreed that auditing or transparency mechanisms were needed, with opportunities for users to provide feedback. Panelists discussed the risk that automated systems might sometimes make mistakes, for example, consenting to a data collection that a user does not want. Dr. Egelman suggested that users currently have such a poor understanding of privacy policies and choices that even an imperfect automated system might result in decisions that more often match a user’s intentions.

Panelists also discussed the need to coordinate multiple versions of policies, and Dr. Schaub suggested that all versions could be derived from a single machine-readable version. Dr. Egelman agreed, and added that machine-readable policies could also help automate the process within a company for keeping privacy policies up to date with actual data practices. Panelists said that machine-readable policies also had an advantage in that they could be presented to the user in a personalized way, and that it would be easier for users to compare multiple policies.

8. Conclusion

Judging from workshop attendance, approximately 225 people in person and 735 remotely via webcast, disclosure evaluation is a topic of broad interest. Researchers presented work from a large number of academic disciplines. Some presenters mentioned they had not been previously aware of the research being done by researchers in other disciplines that was highly relevant to their own work. There may be benefit in examining disclosures through an interdisciplinary lens to take advantage of differing approaches from fields such as marketing, economics, psychology, computer science, communications, and law.

Panelists suggested various ways to improve disclosure design. They recommended using simple, unambiguous language wherever possible, and an organized structure. Some panelists emphasized that disclosures should be designed with the most important or unexpected information first. Some suggested that user studies can help identify what information is most unexpected or important to users, and that layered disclosures can show essential information on a top layer, with links to more detailed information. Some panelists suggested presenting information that shows people why a disclosure may be relevant to them. Some panelists cautioned that if people see the same disclosure repeatedly they may become habituated and ignore it. Others noted that the timing and context of disclosures can have significant impact on disclosure effectiveness. Some panelists suggested that when presenting numerical

information and risks, it may be useful to put numbers in perspective and express probabilities as frequencies.

Panelists also discussed disclosure evaluation as an iterative activity. Some reported conducting initial research studies to provide insight into what beliefs and knowledge people have prior to receiving a disclosure. Some panelists described iterative design and testing to improve disclosures. It was noted that even inexpensive, small-scale studies can provide insights that can help designers improve disclosures. Some panelists suggested that evaluations may need to be repeated over time as technology, public attitudes, or other circumstances change.

Some panelists recommended using multiple methods to evaluate disclosures, including methods that involve asking participants questions, observing participants, and conducting controlled experiments. They recommended that evaluations and metrics be identified clearly before testing begins, and emphasized the importance of appropriate methods for sampling, experiment design, and data analysis. To evaluate attention, some panelists recommended eye tracking, recall, and recognition tests, and noted that self-reports of attention may not be accurate. When evaluating comprehension, panelists said it is important to evaluate whether participants can apply the information from a disclosure, not just whether they understand all of the words. Some suggested that it is also important to understand the rationale behind the decisions participants make in response to a disclosure.

Some panelists discussed approaches for conducting disclosure evaluations inexpensively. They said that recent developments such as crowd sourcing platforms, mouse tracking, and eye tracking via web cams offer opportunities to conduct larger studies more quickly and less expensively than was previously possible.

Finally, looking towards the future, some panelists discussed using machine-readable disclosures that

Workshop team and staff summary contributors

Lorrie Cranor, Chief Technologist
Joseph Calandrino, Office of Technology Research & Investigation
Laura Hosken, Bureau of Economics
Ryan Mehm, Division of Privacy and Identity Protection
Hampton Newsome, Division of Enforcement
Michael Ostheimer, Division of Advertising Practices
Janis Pappalardo, Bureau of Economics