

FTC Putting Disclosures to the Test Workshop
September 15, 2016
Segment 2
Transcript

MICHAEL OSTHEIMER: My name's Michael Ostheimer. Welcome to the Your Attention Please panel, which will be focusing on the first issue that Mike Wogalter mentioned this morning, getting people's attention. And I'd like to introduce our panelists. Our fourth panelist will hopefully be joining us momentarily.

Our panelists are Nathaniel Evans, an assistant professor in the Department of Advertising and Public Relations at the University of Georgia, Mariea Hoy, who is a professor of advertising at the University of Tennessee, David Hyman, who is a professor of law and medicine at the University of Illinois, soon to be at Georgetown, and Rebecca Bakebalo, an information scientist at the RAND Corporation. And I'd like to turn it over to Nate.

NATHANIEL EVANS: Thank you, Michael. And thank you for having me today. So I suppose I will just go ahead and jump into it. First question, why research disclosures in advergaming? Well, first off, several studies actually indicate a low level of disclosure prevalence in advergaming and, in particular, children's advergaming.

Add to this the fact that research indicates when disclosures are present in advergaming, children are not adequately able to understand them and the advertising nature they're in. Our body of research indicates that parents-- and remember that parents are the gatekeepers and socializers-- have a limited understanding of advergaming's advertising nature.

And because advergaming is, in essence, goal oriented, they do require a degree of cognitive effort to play successfully. And this cognitive effort has actually been shown to detract from memory and retention-related outcomes in other gaming contexts, predominantly in the educational psychology literature.

And really when you think about cognitive effort, it can potentially detract from an individual's ability to recognize or understand that advergaming is advertising. Now as suggested by the clear and conspicuous standards, disclosures and their associated design characteristics, seen here, are clear and conspicuous when they help consumers in noticing, reading, and understanding that the communication is an ad.

So in our case, we were interested in how the specific design of disclosures-- one specific design being modality and, in particular, dual modality-- might affect parents' ability to understand that a child-directed advergaming is, in fact, advertising.

So specifically, research suggests that disclosures presented in two different but simultaneous modalities-- so text and audio, in our case-- promotes greater awareness, comprehension, and memory of the information in the disclosure compared to a disclosure presented in a single modality-- so text in our case.

So the question we really wanted to ask was whether this type of disclosure, characteristic, or presentation of dual modality is effective in the context of advergames with parent players. So what did we do? We examined the effect of disclosure modality.

And we compared no modality-- so as the advergame actually appeared in real life-- to single modality and dual modality on parents' persuasion knowledge. And persuasion knowledge, in our areas, is oftentimes referred to as advertising recognition.

So we conducted a between-subjects online experiment hosted by Research Now, which is a market research company. And they hosted and recruited a panel of 202 parents that had at least one child between the ages of 7 and 11. And we asked them to play a Pop Tarts Toastie Turvy advergame. And I don't think this game is available anymore online, but we asked them to play it for three minutes and actually controlled for the amount of time that they played. So no less than three minutes, no more than three minutes.

And they were randomly assigned a one of our three disclosure conditions, being no disclosure as it appeared online, a single modality disclosure which featured a text crawl below the gameplay, and a dual modality disclosure, which in addition to the text crawl also featured a audio voiceover that repeated what was in the text crawl every 10 seconds.

So you see up here, these are how our disclosures were presented within the game. And you can see the single modality disclosure text crawl right here and the dual modality right here. In addition to the text crawl, we just had the voiceover stating what was in the text crawl every 10 seconds.

And it's important to know that because the game was by-and-largely child directed, we had to test for the disclosure language. And we wanted it to fall within the age range of seven and 11, which would be the children who are playing this game. And the grade level came back at a 4.3 reading level.

So what did we find? First off, it should be noted that our modality conditions were actually more effective than no modality. And that's really not a surprise here. But what was a surprise was unlike what we predicted, dual modality was actually less effective than was single modality. And so the superiority of single over dual, I think, has to do with how advergames and combining games with advertising in general tends to result in some cognitive effort, or what we call cognitive load.

And this is especially the case when we compare this type of advertising to more traditional formats. And this really puts pressure on our cognitive systems. And these cognitive systems can be either visual or auditory. And what we think is going on is that the cognitive system that is least burdened is going to be the one that is accessed for determining the advertising nature of the advergame.

So another finding was that we believe that the auditory component in the dual modality disclosure actually competed with some sound aspects of the game, namely some sound effects and in-game sounds. So while the recent Dot Com disclosures recommend that the disclosure

align with the modeling of the environment in which it appears, in our case, placing an audio disclosure within an advergame that has some or actually predominant audio features actual led to a reduction in advertising recognition.

So advertising strategies that include gaming should interest practitioners and businesses. We have 150 million plus gamers in the US right now by some conservative estimates, with a projected \$39 billion to be spent on video games, advergames, and branded apps by 2018.

And I think it was already mentioned earlier today, but really one need to look no further than the recent Pokemon Go trend. And I think it was about two or three days ago I was listening to a radio story about the predominance and rise of e-sports. Specifically, colleges are now offering e-sports classes and even offering e-sports scholarships.

So clearly, given the ubiquity of combining advertising with gaming, I think it's important to ask a couple main questions. So what lessons did we learn, and how can we apply these insights about disclosure and modality to other gaming formats and examples?

So when testing disclosures in other gaming contexts, I think it's important to ask, what is the predominant modality of the game? Is it more textual, like this Frozen advergame with Olaf? Are you required to solve a word puzzle? Or is it more auditory, like this Kellogg's Rice Krispie Treat Cereal Bar Game, where success depends on your ability to match the bar with the music and the words?

And second when testing, I think we have to ask, what cognitive systems are players using? Are they relying on textual or auditory processing centers. And so depending on what systems they use, I think we can garner an understanding of the potential effectiveness either single-- and this could be either auditory or textual single modality disclosures-- or dual modality disclosures.

So keeping in mind our findings-- and I'm going to end on this-- I think it would be prudent to avoid any competing modality when designing and implementing disclosures in these type of environments that actually produce an increased level of cognitive load in general. And I think that's it. Thank you.

[APPLAUSE]

MARIEA GRUBBS HOY: Today, I'd like to talk about if getting consumers to read disclosures is the field of dreams or the impossible dream.

DAVID HYMAN: Use the microphone.

MAREIA GRUBBS HOY: Or rather like a farmer Ray in "Field of Dreams" who we think if we build a disclosure clear and conspicuous, consumers will notice and read it. But as we just heard from Nate, even when you were following FTC guidance, it doesn't always work.

So perhaps consumers are more like Mark, who asks, do you see the baseball man right now? And the advertiser or the client says, of course I do. What? You don't really see them? So even if

we make our disclosures clear and conspicuous, there may be other factors at play. I'd like to tell you about our study where we looked at our consumers reading disclosures.

The context of our study was in prescription drugs, but you can think of other product categories or message strategies such as truth in lending, or negative options or something as simple as quantities are limited. If you ask consumers, do you read the drug risks? They self-report that they do with 40% saying they read half or more. It goes up to 3/4 if it's a drug they're interested in. And 25% agree I always read the fine print.

[LAUGHTER]

So ask yourself, are consumers really reading that much of the disclosure? Do you? Well, maybe the lawyers do, but I know I don't. So what's social desirability got to do with it? Our Western societal norms encourage and expect information seeking. Just think about the popularity of Amazon reviews. So rather than asking if they're reading the disclosure or assuming that they are because we've made it clear and conspicuous, is there a better way to determine if they've really read the disclosure?

So what we did was a three-pronged multi-method approach where we had eye tracking while free viewing a website. And what free viewing a website means, rather than simply saying, go look at the risk disclosure, we said, look at the website like you normally do. And then they completed an online survey, and one of the questions was, how much of the risk information did you read?

And finally, I had in a retroactive think aloud interview with them. And what that means is we pulled the website back up. They put their hand on the mouse. And I said, walk me through how you look at the website. What were you looking for? What were you thinking? And so the website we used in the study-- and thanks to the FDA for providing it-- began with the benefits and having the risks.

It's for a fictitious seasonal allergy drug called Glistele. So if you've got seasonal allergies, please don't go look for this drug. But you can see in the middle there that there are some novel risks.

The people that participated in the study had been diagnosed by a physician that they had seasonal allergy drugs, and have had symptoms in the past 12 months. So they were probably familiar with the medications both prescription and over-the-counter that treated the drug, and probably knew some of the risks. So this was added in to make sure when they did the recall, it wasn't just based on experience, but they'd actually read the risks.

There were two other conditions. One had a signal to the risks with the red bar there, said, look below for more important safety information. And then we had a page with a link to the risks.

Now, our sample size was only 29, so not enough to compare across the three treatments. This is more an exploratory study just to see if the participants were looking at the risks at all. But the risk information was the same regardless of the treatment.

So here are the eye tracking measures that we've already heard some about this morning. Fixation, where did the person look? And that is reflected by that black dot you see up there. They had to have fixated for 2/10 of a second or longer so that they could have encoded the information, and it equates to reading.

Then, there is duration, or how long they look. So I know that they'd look for at least 2/10, but I also know how long was that fixation. And I can mark what's called an area of interest or AOI, which would be our risks. And add up all that total fixation time to see how much time they spend reading the risk information.

And finally, fixation sequence. In what order are they looking? So if you look up there, there's little numbers in the circle. So you might look across the line of text and go, oh, five, six, seven. They're clearly reading across the line. So what did we find?

Well first of all, on the survey, 80% of participants claim to have read half or more of the drug risk disclosure. In reality though, eye tracking data revealed limited to no reading of the drug risk. In fact, on the survey, only 1 of the 29 said, I didn't read any of it. However, 8 of the 29 had absolutely no fixations within the risk information. We had two or three that made a fairly decent effort of reading the risk. But for the most part, there were just a few scattered fixations in the risk information.

So here are a couple of examples. Participant number two, not a single fixation. Scanned all over the place, didn't read a thing.

Participant number three, a lot of fixations, clearly reading the benefits. You can also see a cluster of fixations around the link to the risks. So did they click through? Indeed they did, and they began by reading some of the risks in the first paragraph, glanced down and read some of the risks in the bottom paragraph. But look what they totally omitted, the novel information. So if they thought they were familiar with this drug category, well they would have been a bit deceived.

Then we have participant six, which was more common. Which yeah, they're reading the benefits, but they're just glancing at the risks.

So what were they thinking? They told us that they were reading the risks on the survey. But when we actually looked at the data, they weren't. So why are they seeking out, ignoring or avoiding drug risk information? Well, what were they looking for? In a word, the benefits. What could Glistele do for them? They told us this in the interviews, and it was borne out in the eye tracking data.

So why didn't they read the risk? Well, as we suggested earlier, perhaps because of social desirability. When they were answering the survey, the thought was, oh, I'm supposed to read the risk. I better say that I do. When I did the interviews, this did not surface. But it could be simply that they didn't verbalize it.

How about information avoidance? There were a few of those. I don't want to know about the risks. But for the most part, it wasn't the risks that they were avoiding, but the source. If they wanted to know drug risk information, it wasn't going to be from the advertiser. They'd ask a physician or go to WebMD.

Optimism bias, now this, we saw. I've never had a problem with drug reactions before. It's just an allergy drug. There aren't any risks, really. In other words, they were letting their familiarity or perceived familiarity give them the optimism bias, nothing can happen to me.

Now, this is going to be a particular concern because of those novel risks. They couldn't have known about those risks. And for the most part, they omitted them. In fact, almost entirely omitted them. So they were lying on their own familiarity with the drug category, their experience having the condition. And they didn't avoid the risk, they simply ignored it because they thought they knew it all.

So is disclosure reading the impossible dream? A couple things to keep in mind. First of all, if you ask them, they're going to report a higher level of readership than actual readership. And of course, when we read ads, we're looking for the benefits. What is the product going to do for me? So you can use eye tracking to see if, in fact, are they reading the disclosure or not. And if they aren't, we really need to identify why they aren't reading and come up with some possible ways to address this.

So in our study, what we recommended was if you can, present the drug risks first. When you look at how people went through the website, you can tell they start at the top, and they went through. And of course the first thing they hit was the benefits, and that's what they were looking for. So why not put the risks earlier and have them go through them as they're searching for the benefits?

But more importantly-- of course this could apply to all disclosures, and it's already been mentioned earlier-- create a sense of unfamiliarity if there's novel information. So if in your disclosure you know that you've got the thing that would be atypical for the product category, get that earlier so hopefully when they're reading through, they'll see it. And then also signal that this is new information. This is something you're not used to seeing.

So while we're never going to get all consumers to always read all of our disclosure, by eye tracking, we can identify if they're reading some of it or not. And by interviewing, can find out why they're not reading the disclosure. So maybe, just maybe if we ask, do you see the disclosure right now? We might get more consumer saying, of course I do. Thank you very much.

[APPLAUSE]

DAVID HYMAN: So I'm going to continue the theme that people are complicated entities that are not calculating robots. And you've got a sense of some of the methodologies that we can use to get a handle on that. This is joint work with my regular co-author David Franklyn at the University of San Francisco. We've published a series of studies and articles that evaluate consumer knowledge and understanding of SRP, search results page.

So think of the output when you run a search through Google, or Yahoo or whatever other search engine you want to use, with respect to the layout that is the architecture of the search results page, as well as the ways in which it's labeled. We've also looked at native advertising, which has already been referenced this morning, which is advertising that doesn't really look like advertising.

And we've looked in some detail at the labeling of both SRPs and native advertising, and the extent to which consumers actually are paying attention to that, know what they mean, can make sense of it.

So our basic methodology is considerably different, I think, than you've heard so far this morning, although similar to some of the things that were observed in the very first set of speakers. We've done a series of online surveys. We typically have about 1,000 respondents per study. Mostly US, but we've done a couple of studies in the UK.

And what we do is a diversity of having people conduct simulated searches, and then look at the output and answer questions about it, or we show them static images, and we ask them to identify which of these regions of the page are paid versus unpaid. And then we do a recollection at the back end about whether they remember what they saw in the labeling.

And so we can look at their basic knowledge and their comprehension of what they see. We can also look at and ask overtly about the perceptions of the clear and conspicuous standard and ask them what do particular labels signify. Because there's particularly with native advertising, there's a lot of diversity of labels.

And we exploit changes in the SRP architecture and labeling architecture again, as the way the page is laid out. We can manipulate that. There have also been changes over time. And we can manipulate them ourselves in the surveys and see how it affects the responses that we get.

So let me just show you a couple of images. This is a Google search for Mercedes circa 2010. And I don't think there's a pointer, and the shades don't show up very clearly, which shows you something of the problem with labeling.

But up at the top, there's some paid ads. There's some paid ads on the right. And then there's organic or algorithmic results underneath that, sort of very typical search results page that no one in the world saw prior to around 2000. But after that, everybody becomes increasingly familiar with it.

And there is some diversity. So this is a search for Nikon camera. It's different partly because there's a big hunk of a set of pictures in the middle. That's the Google Shopping region. It's below the regular paid ads, but above the algorithmic or organic results. And then also, you sometimes see it on the right side.

So this is a London fog ad-- I'm sorry, search results for London fog. Looks similar to what you just saw for Nikon camera, except the Google Shopping region is in the right-hand column.

Native ads, I'll just show you a couple of examples as well. This is from Forbes. It's a native ad for Fidelity. It kind of looks like an article about should you accept your employer's pension buyout offer, but it's got the Fidelity logo. It's really hard to see, but it says Forbes brand voice near the logo, and Fidelity Viewpoints Team is listed as the author. This is an ad, but it looks a fair amount like an article in Forbes.

And then here's one from the "New York Times" sponsored by Netflix. That means it's a paid ad for Netflix, but it looks a lot like a regular "New York Times" article on women inmates. This is, of course, for "Orange is the New Black." And then finally, a famous or somewhat infamous example of native advertising that ends up in "The Atlantic." looks pretty much exactly like an article in "The Atlantic," except there's a little yellow thing up there just above the title of the article that says sponsor content.

The other way you can tell it's not really an "Atlantic" article is it says nice things about the Church of Scientology, which I think most people's priors are that would probably not be their default mode. So we've got a series of examples of native ads and a series of examples of SRPs. They use different labels. They're laid out differently. And so you can sort of see how one might approach asking respondents what they understand that they're seeing.

So one obvious way to do it is to just strip the labels out of any context and just ask respondents, do you think this particular choice of words for a label signifies whether the associated content is paid, that is an ad, unpaid content, or you just don't know? So this is just obviously stripped from context. But what do the label words in isolation signal to consumers?

That's one way of testing it. I'm not suggesting it's the only or the best way, but it is one way to compare head-to-head labels without having to deal with confounding variables. And not surprisingly, what you see is that labels that use the words "paid" or "ad" do pretty well. And then there's a hierarchy below that. "Sponsor" does respectably well. Things that use the word "brand" don't do so hot. "Partner" and "Written By" do pathetically bad.

But even with really overt labels like "Paid Ad," there are 4% of people who think it's unpaid content, and 6% of people don't know. This is the, "Who are these people?" question. I had a former colleague who once said, the problem with life is the personnel.

[LAUGHTER]

And you do get a sense of that when you're working doing empirical studies in this space. Similar to that, we in the SRP study in January of 2012, we asked, which of these labels have you seen in your own searches in the last month? So this is in January 2012. And we listed a bunch of different labels. And you can see the percentage of respondents that reported seeing them in the past month or so. And the comments on the right side are somewhat significant.

So sponsored links was the single most popular label. It was only used by Yahoo. I'm sorry. That was last used by Google in 2010. So it would be physically impossible to have actually seen this in the past month. 55% of people, the most popular result, reported seeing it. And these are not click-throughs. We have controls to deal with that problem.

Sponsored results is actually in use in the real world, but only by Yahoo, which had a very small share of the search market, both in the world and in our study, so very high numbers. Ads, which is the label that was actually in use by Google at the time, 46% of people reported seeing it. So third on the list actually in use.

Commercial ads, which had never been used-- we just came up with it because we thought it would be a good control-- 33% of people reported seeing it in the past month. And then 22% of people just weren't paying attention to any labels whatsoever.

So we exploit variation in changes in labeling, and this also highlights the importance of including controls in your study. What about native ads? Well, we just asked, we showed people a bunch of native ads and said, is this an ad or not? And we also used unpaid content and regular ads as controls. The blue bar on the bottom shows you the percentage of people that thought these were ads or paid ads.

And the results are not really the thing I want to highlight here. The point is to understand the methodology. You show people content and you ask them, do you think this is an ad, or do you think it's news or editorial content? But if you care about the results, on the right bar 81% of people presented with our regular ads, and we tested a bunch of them, thought they were ads. But only 37% of people presented with native ads thought they were ads. That's averaged across everything we tested.

There's a little bit of variance there, but people obviously do much better with regular ads than they do with native ads using the methodology that we did. The other thing you can do is tweak or modify the labeling of tested native ads and see if you can actually make a difference. So this, you can think of this as an AB testing approach, where a is the control and b is the modification.

So this is something I showed you already, the Forbes Fidelity content with the labels that I mentioned before. What we did is we added a big honking bar right in the middle of that said "paid ad" in bold black letters and a gray background. And then we tested across sample, do you think this is an ad? And you can see the results here. We move from 40% for this particular ad for people who saw it in its original form to 56% of people who thought it was a native ad.

So again, the point is to understand the methodological strategy. You modify and see if it makes a difference rather than just assuming, oh, we'll slap something on it and people will notice and they'll pay attention to it. Very consistent with the themes we've already heard today. And then let me close just by leaving up the papers that we have in this space. Thank you very much.
[APPLAUSE]

REBECCA BALEBAKO: Hello. My name's Rebecca Balebako. I am going to talk today about a paper called "The Impact of Timing on the Saliency of Smartphone App Privacy Notices." So we're moving away from advertising, and now talking about privacy notices. I had a really great collection of co-authors on this paper.

So what makes a privacy notice effective? Well, definitely that notice should have information that people care about, things that they want to know on it. But also, that privacy notice should

be salient. People should turn their attention to it and pay attention to it. And that's what we're talking about in this panel. So one way to measure salience is memory. Do people actually remember the notice sometimes afterwards? And that's what I used as my measure in this study.

At the time that I did this study, this is what smartphone permission notices look like. So on Android, when you were installing the app, you got a little notice that popped up. And there was plenty of research showing that not only did people not understand these notices, but if they did notice them, they didn't actually understand what was in the notice.

Also at the time, this is what an iPhone notice looked like. So this was a just in time notice that would show up while you were using the app and asked for the permission right at that time.

So there had been some research on the content of the notices or some work on what the notices looked like, but I was really curious, does the timing of these notices matter? One's being shown in the App Store. The other one's being shown while you use it. Does it matter? So we brainstormed a bunch of different timing options.

So when, if you were to install a smartphone app, when would be some possible times to show a privacy notice? Well, you could show them in the App Store. So before you even click on Install and make a decision to download that app, you could show it during install, so part of the install process, which is what Android did.

You could show it before app use. So its already on the phone. The person's opening the app, and they might get it with this flash screen, or right after this flash screen. Could be during app use, so like a just in time notification, you're using the app and you get a notice. Or it could be after app use.

A lot of people ask, well, what's the benefit of getting a notice after you've used it? But you can consider this for a case where you might want to give someone a summary, like Angry Birds collected your location 1,000 times in the past day. So that might be an after app notice. That's when it might be reasonable.

We did not actually test during install, because we did do real phones and we were not able to actually hack the Android install process. So we tested the other four conditions, in App Store, before, during and after app use.

So I'm going to tell you the result, and then I'm going to talk about the method. So we didn't find out which timing was best, but we did find out which one was worst. In the App Store was the worst time to show an app privacy notice. So let me tell you a little bit about how we got there.

Well first of all, we had to design an app. And so we developed a pretty simple app quiz. It was 11 questions on African-American scientists and inventors. And it was really pretty simple. It was multiple choice. This is what it looked like.

In fact, it was so simple that even though we designed this app to be available on every platform, Apple rejected it and said, this is just like a web quiz. This isn't good enough for our App Store.

So I tried to make the case, hey, this benefits humanity. This is research. They didn't buy it, so all of my field test experiments are based on Android.

And so this is the privacy notice. And you'll see it doesn't look like any of the previous notices that I showed you. So we shouldn't have to deal with acculturation, I mean, people being too accustomed to the notice. And we were really careful about trying to design a notice that people would care about, that they wanted to remember, but wasn't so scary that they would immediately stop the study.

So we followed the NTIA code of contact on mobile app privacy notices. And there were two pieces of information we had to choose, what information is being shared and with whom it is shared. So we tried to based on some MTurk study, some previous work, we chose two that were sort of middle of the line. So browser history is being shared and it's being shared with ad networks.

So that's what the privacy notice looked like. It is standard across all conditions. So we weren't testing the content or what the privacy notice looked like. We were just testing the timing.

So we did both a web survey and a field experiment, and each one has its benefits. The web survey, we MTurk participants, and they played a virtual app online.

So they were on their web browser, and then there was a little black box. And we said, pretend this is your smartphone. And they had to go through the process of virtually installing the app and playing the game, whereas the field experiment, we had real people actually installing the app on their phones in the wild. So we weren't sitting with them. We weren't watching them. We asked them to do this and to play the game that we gave them.

For the field study participant recruitment, I don't have a lot of time to go into details, but I am going to point out that we polled participants from five different groups. So it was a lot of work to recruit them. They were a little more expensive, but there was definitely a lot of benefit in the realism of a field experiment.

So the participants completed the following steps. First, they completed a consent form and some demographic questions. So the point of this is if people dropped out, we wanted to know what kind of people were dropping out. We didn't find any significant differences between demographic in dropping out.

Then, we asked them to install and play the app, so at which point they were randomly assigned to a timing condition. They would have seen the notice during the step two.

In step three, they experience a distract or a delay. So in the web survey, we asked them a number of questions about their privacy preferences. In the field experiment, we waited 24 hours after they finished playing the app and automatically sent them an email saying, please come back and answer some questions.

And so these questions included memory questions, so specifically recognition multiple choice questions about the app. What was the color of the background of the app? What were some of the inventors you were asked about? What was on the privacy notice? And so that's what we used as our main outcome.

And then we showed them the notice, and we said, hey, this notice isn't true. We didn't actually collect this information. But can you look at this notice and tell you whether it has stuff that you would want to know and you would want to remember? Most participants said, yes, this notice has information that I would want to know and I would want to remember.

So the final results. So one thing I should point out here is these are the different conditions. We included a condition like a control condition called not shown. So this means people didn't see the notice at all. We didn't show them a notice. So we anticipated that everyone should do better who saw the notice than the people who didn't. And you see in the web survey that this is true, they did do better.

However, the color means, was that statistically significant? So we compared all of these conditions to not shown. And unfortunately, there was no statistical significance between the App Store and not shown. So this is pretty sad.

Before use, during use, after use all better than not seeing the notice at all. All statistically significant. But the between comparison between these conditions before, during, and after use, there was no statistical difference between those, which is why I say I can't tell you which one is best.

The other thing I should point out to you is that the total rate of-- it's a typo on the slide that it says "recall." It was actually "recognition." The total rate of recognition was pretty low. The axis does not go up to 100%. It goes up to 45%. So even the best condition was getting around 40-some percent. And the total overall across all conditions was 25%.

These are the results for the field study. So similar results. App Store looks better. But the difference between the App Store and not shown was not statistically significant. And the difference between conditions before use, during use and after use, that was not statistically significant within them, but they were all better than not shown. Yes, so they were all better than not shown.

So one thing that's really interesting about doing both the field study and the web survey is that the differences between the conditions all went in the same direction. So that provides a lot of-- I shouldn't say a lot of. It hints that there is some external validity with doing a web survey, which is great because a lot of the previous research on smartphone app notices had been done on the web, and not necessarily out with the field experiments.

So that concludes my study. Yeah, happy to answer questions. And there's the title of the paper. Thank you.

[APPLAUSE]

MICHAEL OSTHEIMER: Thank you, panelists. Could we turn the mics on down here? Great. Thank you. Now, I'm going to ask the panelists some questions. And we probably want a time for questions from the audience, but we'll see. My first question is, why did you choose the methodology you used? Maria, would you like to answer that question?

MARIEA GRUBBS: Well, I think one thing is that eye tracking just gives you that objective assessment people are looking. And so rather than going, are consumers really reading risks, and of course it was a drug risk, but it could be disclosures of any kind, you get an objective measure with eye tracking that you don't get anywhere else.

But we have that caveat that eye tracking is simply one tool in your tool kit. And what I learned most from this study is the importance of the multi-method approach, that we really benefited by having eye tracking data combined with survey data combined with the qualitative interviewing afterwards.

MICHAEL OSTHEIMER: All right, David?

DAVID HYMAN: Yeah, I didn't have an eye tracking machine.

MAREIA GRUBBS HOY: You want one?

DAVID HYMAN: Yeah, sure. Is it free? No, actually this grew out of some other work that we had done. And so we were familiar with the methodologies we were using. We, as I think good practice, sat back and said, well, is this the best way to get at the problem that we're interested in?

And I really want to echo the observation about the importance of multi-method modalities in studying this, all their trade-offs with all methodological choices. And when you can triangulate, you should have a much higher degree of confidence, particularly across studies than if you just use one method and declare that it's the best one, because that's the only one you know how to do.

MICHAEL OSTHEIMER: Great. Anyone want to add to that? My next question is how costly was the methodology you used, and where there are any cost-related trade-offs? Nate?

NATHANIEL EVANS: Yeah. So the panel was fairly expensive. I think it was over \$3,500. Now, we considered doing a laboratory experiment-- but actually, this is out of my dissertation-- so timeliness was key. And that's one of the benefits of using a panel.

So really, it comes down to prioritizing expediency for me. Another cost-related trade-off was the consideration to design a new game. And that is vastly expensive in speaking with somebody at the Label Interactive up in Connecticut. Well, he gave me a low-ball estimate of \$5,000. And as a doc student, that's pretty much out of the range. So really, it was an existing game, and that was a cost-related trade-off.

And when you choose an existing game, you bring in some issues as well, mostly familiarity with the potential brand. So that's just one of the trade-offs you have to be aware of.

MICHAEL OSTHEIMER: All right, David?

DAVID HYMAN: Yeah, so money matters, particularly if you work for a state university. My co-author is very effective at raising money, so that helped. Your choice of panel makes a difference. So MTurk is going to have a very different cost structure than if you use a commercial survey organization. And there's price variation among them as well.

How big your sample is, you multiply it by the dollars per person, and then there are a variety of fixed costs, depending on what you can do yourself. So if you're using existing SRPs or native ads, it's going to be cheaper than if you're trying to mock them up yourself. Some people I suspect outsource the statistical analysis, which could add an additional cost. So you just have to think about what your budget is and how you want to spend it.

MICHAEL OSTHEIMER: Thank you.

MAREIA GRUBBS HOY: Oh, I could speak a little bit to eye tracking. We did not outsource. It our college got an eye tracker back in fall of 2011. And my understanding is that that machine cost about \$30,000. And then we had to have a lab tech to run it. I don't run the equipment. They do it for me. So I'm not sure how much it cost to outsource it, but it is not inexpensive to acquire it at a university.

MICHAEL OSTHEIMER: What are the advantages and disadvantages of an online study versus a field study? Rebecca, would like to take that one?

REBECCA BALEBAKO: Yeah. So a field study, you get people actually installing an app on their own phone. And in this case, particularly for something like this, it's different than just doing it online and having this virtual environment. However, with the MTurk study, we were able to get more participants. They were cheaper, it happened quicker, and we were able to get participants who used every major smartphone platform.

And we also had a number of participants who didn't own a smartphone. So you may wonder why this would be valid, but I'm assuming at some point in their lives, they may actually get a smartphone. So it's very interesting to see what their own reactions to these notices are.

Also, the MTurk study, we had from participants from 49 states, whereas the field study, we didn't. So the online studies, you can get a broad swath of participants, pretty good diversity, something where you're actually getting people to do something on their own device. It's a little bit harder to recruit for, but it's more realistic.

MICHAEL OSTHEIMER: Did you notice any systematic differences between the two modalities?

REBECCA BALEBAKO: So apparently it's not kosher, because it's different test to put them into the same statistical pool and run some sort of test to see if they're different, but I did it anyway, and I didn't see any differences.

MICHAEL OSTHEIMER: What, if anything, would you avoid doing in testing the noticeability of disclosures? Rebecca?

REBECCA BALEBAKO: What would I avoid doing in testing the-- so one thing that I particularly for a privacy notice, you don't want to inform your participants that it's a study about privacy notices. And that actually happened by accident with one of our participant pools, where an email went out saying, hey, come participate in a privacy notice study. And I said, wait, wait, take all those emails back. We can't have those participants.

But also, because it was we were telling them that we were collecting this data, but we didn't, we had to be very careful about the disclosure at the end. So you have to think carefully about what all you're telling them about what you're going to expect them to notice and look at before you actually start the study.

MICHAEL OSTHEIMER: David?

DAVID HYMAN: Yeah. So on the things you don't want to do, well, one of the things you don't want to do is run a study without controls. You should always ask where's the control when you're designing the study or reading somebody else's study. Because if they don't have a control, you have no idea whether the results are valid, let alone whether the magnitude is remotely right that's being claimed.

So I showed some examples of using labels that were no longer in use and made-up labels, or fanciful labels. Those are useful controls.

The other thing you don't want to do is assume everybody's behaving in good faith in filling out your survey. You already heard about the click-through problem. And so if you're getting results that you expect but you haven't assessed the click-through problem, you don't know whether you should actually believe it.

So one obvious way of doing it is ask is 5 plus 7 equal 12? Or, what is 5 plus 7? And then you give people five answers, and one of them is 12. If you've omitted the 12, you've made another mistake, which you should not do. But you won't deal with the person who happens to be clicking through on the right if it's C. If they click every one C, you'll still have some click-throughs that remain in your pool.

The other thing consistent with that is people, even if they pass the sort of click-through test for-- and you can do that various ways-- you also want to worry about people who take either way too little time or way too long time.

And there are some rules of thumb here, but somebody who has the study open for an hour, you probably should not think about including them. And somebody who completes the study in two

minutes that everybody else takes five minutes for, that's probably at the low end. So those are the sorts of things I would think about.

The last thing not to do is write the question, and throw it into the survey and assume that you've actually captured what you want to capture. You should do a serial process of looking at your questions over and over again. Asking, is this actually capturing what I want to capture? And that's an area where co-authors can be helpful, because they don't read the question with the same assumptions that you bring to it, because you drafted it.

So I certainly agree that too many co-authors can spoil the soup, but a few co-authors can actually improve your questions.

MICHAEL OSTHEIMER: Thank you. Knowing what you know now, are there any methodological pitfalls you would avoid in testing disclosures? Nate?

NATHANIEL EVANS: Yeah, one is timing. Timing is very important depending on your stimuli. And what I mean by timing is the amount of exposure somebody's going to interact with your stimuli. And so some colleagues and I have approached this differently, specifically with native, we've let it naturally vary, so as the participant to read this as you would in a normal everyday situation.

However, with the advergaming study, we wanted to control for three minutes. So we asked them, play for three minutes specifically, and research now cut it off when they're at three minutes, because they have more opportunities to be exposed to the disclosure, which could influence the result.

Another thing that I would avoid doing is using an existing game. Unless you have access to the source code, it's very difficult to actually design a disclosure within the environment. In fact, it's impossible. And so when we designed this, the computer engineer, which I have no knowledge of, actually said, OK, we'll create a window overlay, which I still don't exactly know what that is, and then put the disclosure on top of that window overlay right below the game.

So there are these design aspects that I wish I knew going in. And having money makes a world of difference to actually design your own game.

MICHAEL OSTHEIMER: Do you have anything to add on that question, David?

DAVID HYMAN: No.

MICHAEL OSTHEIMER: How useful and reliable is eye tracking for testing the noticeability of disclosures?

MAREIA GRUBBS HOY: Well in terms of reliability, I'm assuming the technology of their equipment has been out there a long time. So is it working accurately? Correct.

In terms of validity, the reason that eye tracking is supposed to work is what's called the mind-eye hypothesis. In other words, you're probably thinking about what you're looking at. But we've all had those moments we've been looking at something and our mind was a million miles away. I probably did not have that happen at all. In fact, I was looking through a one-way mirror, and I did not see that occurring during the experiment. Of course, it's an experiment and people are behaving differently.

There are some things to keep in mind when you're looking at the eye tracking data. For example, we got all excited-- oh, look at the fixations. It shows people are reading. But on one hand, too many fixations, especially if you look at the chronology of it and it's back and forth and back and forth, can show people are being confused by it, not that they read it. Remember, you're just simply seeing, did they look, not did they comprehend. You're not assessing reading comprehension, just if they looked.

There were some other issues that we ran into, and this could be an artifact of the age of the equipment. It's now five years old. And that involves being able to actually distinguish between the pupil and the iris, which is how the eye tracker is capturing movement. I lost three participants because we could never calibrate them because their eyes were so dark brown that we could not distinguish the pupil from the iris.

Another thing that we had happen is that you have to be perfectly still. So say, get comfortable, and then they calibrate your eyes. This is going on. And through the one-way mirror, and my tech looks up at me, and just shakes their head, and was like, OK, lost another one. So I lost six people's data because of something dealing with actually collecting the information. So those are some things to keep in mind.

And I came across a Thoreau quote yesterday in my Twitter feed, the source of all knowledge. "It's not what you look at that matters. It's what you see." So eye tracking measures what they look at, but doesn't tap into, what are you really seeing and what are you thinking about as you see? So once again, the exhortation, use it in conjunction with other methods.

MICHAEL OSTHEIMER: How do you make the environment of a study realistic? How important is it to have a realistic and how important is it to have a realistic environment for a study evaluating attention? David?

DAVID HYMAN: Well, I think it depends on what you're trying to do. Some things you absolutely need to and want to study in the environment that it is out there, because you want to actually see in the context in which ordinary consumers experience it, what they perceive, what they understand and how they behave.

But the difficulty, and this is the sort of reductionist question, is how many things do you want to abstract away in order to try and get some insight into a simpler question? So all hypotheses are wrong, but some are more useful than others in testing what it is that you want to evaluate. And so we used a variety of methods.

So if you look at the things I've shown you, abstracting the labels away from the ads and SRPs in which they appeared strips out all of the richness. On the other hand, I'm not particularly interested in testing the richness in that setting. I just want to ask, if you see this label, what does it signal to you? And if it does poorly, it doesn't matter if people actually notice it. It's going to do poorly.

But on the other hand, if you then look at how people actually interact with the label in the context, you get more information. So I was trying to get at different things, and that's why I use both a reductionist approach and a much richer approach in the same study.

REBECCA BALEBAKO: And I think it's also very hard to get a very realistic environment. Because if you're being ethical, you're telling them that they're part of a study. And so their behavior is going to be a little bit different. They might trust the researchers to not give them an app full of harmful privacy risk things, so there may be more trust there.

And they may be more likely to complete a study or go all the way through it as opposed to maybe not, in my case, installing an app because it's a different environment in the study. So it's important to try and be realistic, but it's also, I think, somewhat impossible.

MICHAEL OSTHEIMER: Do you have any expectations about how disclosure and noticeability in a lab study is as opposed to how it works in real life?

MAREIA GRUBBS HOY: Well if anything, they probably are more enhanced. You're probably getting an inflated readability in a lab because they are focused on the task. They're there to do the job. And in my case, they were getting a Walmart gift card for their participation. So in the real world, we're probably not paying nearly as much attention as we are in a lab situation.

MICHAEL OSTHEIMER: Rebecca, anything to add?

REBECCA BALEBAKO: So I think in mine as I mentioned earlier, the participants might've trusted us, so they might have actually paid less attention to the notice than they would in the real world, but it's not really clear.

MICHAEL OSTHEIMER: David?

DAVID HYMAN: So the other complication is the sample population versus the general population. So if you're testing undergraduate psychology majors at public and private institutions whose parents are spending a lot of money to send them there, you may get quite different results than with an elderly population, than with a working class population. And I will tell you that psych students are different than law students because psych students want to make their professors happy, and law students don't.

[LAUGHTER]

And that will affect their behavior in any study you're going to run.

MICHAEL OSTHEIMER: Great. Thank you, panelists. Do you have any questions for each other, or anything else to add? Great. Thank you very much.

[APPLAUSE]

NATHANIEL EVANS: Thank you.

RYAN MEHM: Good afternoon. My name is Ryan Mehm, and I'm an attorney in the FTC's division of Privacy and Identity Protection. Disclosure comprehension is the focus of this panel. Well, our last panel looked at whether and when people are noticing, reading or paying attention to disclosures. This session will examine studies that evaluate whether people understand the information conveyed in disclosures.

We're delighted to have with us today four speakers who've spent a lot of time focusing on this issue. To my right is Dan Goldstein of Microsoft Research lab. To Dan's right is Betsy Howlett, a professor of marketing at the University of Arkansas Fayetteville. To Betsy's right is Susan Kleimann president of the Kleimann Communication Group. And to Susan's right is Joel Reidenberg, professor of law at Fordham University.

Each presenter, like the last panel, will have 10 minutes. And after all the presentations conclude, we will have a moderated discussion. And with that, I will turn it over to Dan, who will lead us off. Thanks very much.

DANIEL GOLDSTEIN: So we're going to be talking about comprehension. I've been specializing in comprehending difficult to understand numbers, like probabilities, percentages, investment returns, very large, very small numbers. And we're going to look at some techniques, evidence-based techniques that have been used to get people to understand these numbers and talk about the ways that we've been able to test them with web panels and other techniques that have been used recently.

So to convince yourself this is a issue, take for instance this question that was asked to a sample of 1,000 Americans. How many times larger is a trillion than a million? Take your own private guess. 1,000 Americans are asked. 12% of them said they didn't know. But of those who said they did know, about 3/4 of them actually didn't know. If you want to check your own comprehension, it's 1 million times. So it's a pretty uniform distribution across the responses, though.

Or how about how big is 100 million acres? Is it as big as the smallest state, Rhode Island, Connecticut, South Carolina, Illinois, Utah, California? For a little fun, let's raise our hand and vote. Who thinks it's as big as California? You're all wrong. It's as big as California. So these numbers are hard to understand.

We are testing a very simple technique called a perspective sentence that we define formally. It's one of these sentences you sometimes read that says, to put this in perspective, 100 million acres is about the size of California, for instance. We wanted to be concrete about what we meant by perspective, so we made templates, kind of like Mad Libs, where you can create perspectives.

And we had crowd workers create perspectives for numbers that appeared in the "New York Times."

So to give you a flavor of some of these, \$100 million is about equal to percent of the blank of blank. Somebody could say it's about 33% of the net worth of Jennifer Lopez, which happens to be true. Or the GDP of the island nation of Kiribati, which also happens to be true. So those are what perspective sentences are.

And to evaluate how well people comprehended text with and without perspective sentences, we took 12 quotes from the "New York Times" like this one, "This storm killed thousands of people, left one million homeless and destroyed was left of a declining banana industry, once the country's lifeblood," and just asked people to write their own perspective sentences.

Ordinary people on Mechanical Turk came up with very good perspective sentences voted on by the others, such as 1 million people is about 12% of the population of Honduras. So that's the perspective. Now, the question is, how do we test whether this improves comprehension?

So what we did is a series of studies in which we had people read news articles as they appeared in the "New York Times" or as they appeared in the "New York Times" with this one little sentence after. So with the one sentence after you'd read, "It left a million people homeless. To put this in perspective, that's about 12% of the population of Honduras."

By the way, if you think about that, 12% of the population homeless, that puts the whole article in a different light, doesn't it? Imagine 12% of the population going homeless from one storm.

So how do we test this? Comprehension is something we can assess directly, but we can get at it through different measures. So recall, recognition, they're not just good for awareness. They're good for comprehension. If you read a number and don't comprehend it at all, it's very unlikely that you can recall it a few minutes later.

So we had people read the articles either with or without perspective sentences. Five minutes later, we said, oh, by the way, how many people were left homeless? The group who saw the 12% number was much more likely to recall that it was a million than the group who didn't.

Now, it's not always the case that you're going to recall exactly what's in a disclosure. You probably recall very little that's exactly in the disclosure, but you can get close. So we also looked at estimation. So what fraction of people are off by one order of magnitude? What fraction of people are off by two orders of magnitude or less, and so on?

So we can look at recall, educated guesses, both of these are improved by adding perspective senses. You can look at error detection. Give people two versions of a text, one that has the right number in it, like the APR on this credit card is 5%, another has a wrong number on it, it's 25%. See how often they detect the incorrect number.

And finally, long-term recall, we found that just adding these sentences like 12% of the population, helped people three months later recall the numbers that they had read in the "New

York Times." And they didn't just read one quote and were tested. They read six quotes and were tested. So it's impressive small technique you can add to disclosures, and it's in this paper Barrio Goldstein and Hofman that you can find on my website.

There is potential to use this in labeling. So for instance, you can take nutrition labels, which might just say, a soda has 250 calories. You can put this in perspective by saying it's 11% of your daily calories. It's about a third of a typical meal that you might be having. You can use exercise equivalence, which has been popular recently. 250 calories is about 50 minutes of walking.

Or as Ryan pointed out to me, you could express it in cups of shredded lettuce-- 31 cups of shredded lettuce. Shredded lettuce makes every food seem very caloric.

[LAUGHTER]

Why not do this with things like front-end load on mutual funds? Instead of saying, a front-end load of 5.25%, which could go in one eye and out the other, you could say, if you invested 50k in this fund, you would pay \$2,625 of that as a fee. Empirical research is needed here, but what we've done so far in general numerical comprehension in the news suggests that it could work.

Another simple technique that can be leveraged is just representing risks as frequencies. So replacing 60% with 6/10 seems like it wouldn't help, but another comprehension check that people have used is they give people these difficult to solve probability problems, and they just replace the probabilities with raw frequencies. So they say 6/10 instead of 60% or 0.6. And then what do they find? The proportion of people who solve the problem correctly goes up. It's a very simple change.

If you're still skeptical and you doubt that it works, consider this. What is 0.01% in terms of frequencies? That's one in what? It takes a while to figure it out, right? In some segment of the population, the majority of people cannot answer that question correctly. It's 1 in 10,000. So that tends to work. And you can get at it by giving people problems, looking at the solutions to those problems, just like I gave two examples there.

Icon arrays are the visual-- we're moving into the visual now-- analog of what I've just talked about. They show frequencies visually. So here, you can take a complicated problem like showing that there's a certain incidence rate of a disease. A certain number of women, say, will test positive for breast cancer. In this illustration, six will test positive. One will test positive and actually have the disease. And now, you can ask people, what's the likelihood that somebody who tests positive has the disease?

They find that by introducing a simple visual like this, you can go from the majority of people getting the wrong answer to the majority of people getting the right answer. And the effect really holds. People trained with these icon arrays can remember how to figure out the problem six months later.

So moving from stated representations to visual representations, I think the future is going to be these simulations that have been very promising when showing investment returns. So there's a

lot to convey in an investment return, because it's not just one number. It's risk plus variance and sometimes an unusual distribution of numbers. So what is my investment going to return with what likelihood?

It has been shown that just simulating draws from that distribution, saying, OK, imagine you held this investment for 10 years. Here's 10 possible outcomes from that investment. That actually helps people estimate the likelihood that they're going to lose money, the average return of the investment and so on. Surprisingly, it can even do better than telling people what the average return on the investment is and telling people what the standard deviation of the investment is.

And this is a graphic from a tool I created. I don't know if you can start. This is actually a video. I don't know if you can get it to animate, but it should. Can we go back a slide and click on the video and get it to animate? Yeah. Can you get this to animate? If I click the button, it's just going to advance to the next slide. You can't animate. OK, so if it were going to animate-- there it goes.

This tool, you can ignore the first part. In the first part, people are basically putting probability into a distribution, so they're creating an investment that has 100 possible outcomes. And then when they click Done, they're going to simulate a return from that investment. So you're going to see 99 of the markers disappear, and just one left standing. So pretty soon, they're going to finish creating the investment.

And this is all the part we can ignore, but this is the simulation aspect right here. When you click Done, you see what it's like to take an investment and have it pay out. So what we have on the vertical axis is wealth. And where the one person is left standing at the end, that's how much well that person ends up with at retirement.

So by doing this over and over again, you get the feeling for what this investment could return on average, what's the worst it's likely to return, what's the best it's likely to return and so on, in a way that is as easily experienced as observing natural frequencies in the world, and without writing probabilities. Which that has advantages for less numeric populations. So I'm out of time. Thank you.

RYAN MEHM: Thank you.

[APPLAUSE]

RYAN MEHM: Just hit the green button to advance.

ELIZABETH HOWLETT: Thank you. Well, it's great to be here. Thank you very much. It's so exciting to see so many people interested in learning and discussing information about consumer disclosures. What I'm going to be talking about is work that my colleagues and I have done over a number of years. Most of what we've done has focused on nutrition issues.

And in particular, this is a project about front of package disclosures. And it was supported by some work from the Healthy Eating Research.

So if you have been to the grocery store lately or if you've been online ordering groceries from Amazon, you will know that the front of a package is a very, very crowded place. It's very all kinds of information about labels, and claims, and fat-free, and trans fat-free, and real. All kinds of information on that label.

So it's difficult for consumers to really be able to sort through a lot of this information to be able to choose the item that is best going to meet their needs. So from my perspective as someone who's interested in facilitating consumers' health and welfare, we want to have them pay attention to the important things, which is the nutrition facts panel. What's in the ingredients? And so here's two example of the new nutrition facts panel.

And also in terms of nutrition, but you see in comparison to all the beautiful bells and whistles that you have on many different kind of packages, you kind of lose your way.

So why did we really start looking at this? One of the issues that parents were asked is like what is the number one problem that you're worried about? And it wasn't smoking. It wasn't drugs. It was actually childhood obesity. So in general, what this tells us is that consumers really do want to make more healthful choices for their families, and so that's a really big issue.

So what can we do as academics to inform the decision-making processes so that consumers can comprehend what's on the package? So we look at this from the perspective of scholars. But also, I'm at a business college. And so how can we help retailers, how can we help manufacturers basically provide the information that is going to be most easily understood and used by those consumers?

And then of course, here we are at the FTC. People from the FTC and the FDA, of course, are very interested as well.

So the basic processing and our basic question was, how does information disclosures at the shelf level influence consumers' understanding? And one of the main concepts that we really focused on is processing fluency. And so processing fluency is basically this understanding of that information.

And so you've got two types. You have this perceptual, so we do things that you saw this morning, red lines around things, things are highlighted yellow and so forth. And then you can also have this conceptual. And conceptual's really, what does it mean?

And so if you look at the three icons here, you have obviously that's caution, a little bit more in terms of its a danger to have your hand squished, or a suffocation hazard. So you want to make the conceptual fluency, you want to make consumers be able to understand what that is.

And so we looked at this within the context of comparative versus noncomparative processing. So a lot of the research that goes on in our academic world has really focused on the

noncomparative. What do you think about this one app? What do you think about this one product? But in the real world, a lot of consumers are actually standing in front of the shelf, or they're online and they're doing different kinds of comparisons between different options that they can order from Amazon.

So when you look at the different kinds of front of package labeling systems, you see we have all these different kinds of information systems out there. We have the nutrition keys, which is something that the Grocery Manufacturers Association is using, different kinds of things, for example. 0 trans fat, 450 calories, and these are our objective kinds of labeling systems.

And then, we also have more evaluative, which is basically it's going to inform consumers about that. For example the Great For You, which is an icon that Walmart uses on their products. Or Smart Choices, or the Guiding Stars, telling three stars, this is good. It's something with three stars, and that's good for you.

So what we looked at is we looked at comparing these different kinds of icons, the objective icon versus the evaluative icon in both a noncomparative and a comparative processing mode. So when you're looking at a nutrition keys where you're having calories, and grams of sodium and so forth, that's objective information, and it's very, very diagnostic when you're looking at a noncomparative situation.

But when you're doing something in a comparison, and you can see a product with two stars versus a product with only one star, perhaps no stars, that's more helpful when you're doing multiple comparisons.

So we did different scale endings in terms of perceived conceptual fluency, something like, given the information on the front of the package, it's easy for me to determine how healthy the product is. The information on the front of the product package, it's clear whether the product is high or low in terms of nutrition and so forth, so perceived healthfulness.

And then of course, as a marketing person, purchase intentions. Are you intending to purchase this product? And so the first thing we did as a number of our other researchers this morning have talked about, this was a Mechanical Turk study, and we did it online. And in one case, we had an objective option present, and then we also had an interpretive.

And we did a simple noncomparative. So we said, well, how healthful is this cereal? And you see that in a noncomparative situation, when you're showing that nutrition keys, the consumers are doing just as well in being able to determine or comprehend the nutritiousness of that product when the objective information is presented alone as when we also throw in an evaluative.

The evaluative is not helping at all. If there's no information on the package at all, and you put an evaluative, that's going to help. So that was the first study.

The second study is we wanted to do something that was more realistic, something that was going to more closely replicate what consumers actually do. And so we did this in our lab. And

we have a retail lab that provides the opportunity to do very realistic kind of research. So the consumers actually have an actual basket and they're asked to go shop for products.

And in this situation, we use something that was very similar to the Guiding Stars. We had 0, 1, 2 or 3 stars. And then in the other condition, we had something that was close to the nutrition keys. And here's a couple of pictures of our lab.

But what we found in this situation is the reverse. So when you have a comparative processing situation when you have people standing in front of the shelf, and they're looking at different options of whether it's-- we did it for a number of different product categories. We did it for cereal. We did it for macaroni and cheese. We did it for soups. And we did it for granola bars.

Basically, what we find is when you show an evaluative icon, it's much easier for consumers to determine which is the most healthy choice. And they looked at this and said, oh yeah, I know which one is the most healthy choice. And then adding the more objective information doesn't help. It doesn't help you make a better choice.

And so these general pattern of results were replicated in terms of the perceived healthfulness. This is really going to help you determine which is the healthy icon when you have the evaluative option. And in this case, again, if you have that comparative icon, it's going to help consumers be able to say, yes, which is the healthiest and purchase intention.

And so if you're looking at what consumers are looking to purchase, when you have an evaluative icon, it's going to increase consumers' desire or willingness to purchase the most healthy icon. And in terms if you go to the more objective icon, you're not having very much impact on that product.

So that's just sort of a conceptual model of what we're talking about, is that one size does not fit all. There really has to be a recognition and an understanding of the type of processing conditions that consumers are going to be participating in. That is, whether they're tending to look at the product in isolation or whether they're going to be comparing the product to different options that are available either online or at the shelf level.

So that's basically the important points, the kind of processing mode is really going to influence your purchase intentions and your conceptual fluency. So thank you so much and [INAUDIBLE].

[APPLAUSE]

My dog is not that well trained.

RYAN MEHM: Thank you, Betsy. Next up is Susan.

SUSAN KLEIMANN: Well, it's great to be here. Thank you very much for inviting me. I'm going to be talking from a slightly different perspective, like Ilana. We are a research and design

firm. And when we go in to work with our clients, we have a problem to solve-- often, a very tight time frame.

In this case, we were working with a project for the Consumer Financial Protection Bureau, in which they were mandated by the Dodd-Frank Act to produce a combined disclosure for mortgages. And they were to produce that within a year of their stand up. That doesn't give you much time when you consider that there also has to be a competitive bidding process. So we didn't actually get to get into it until about six months after they stood up.

What I'm going to talk about is that these were our goals of the project. So we had a comprehension goal, we had a comparison goal and we had a choice goal. And I'm going to take you relatively quickly in 10 minutes through basically some of our design choices, how we set our study up, and then how we did some of our analysis and how we were able to verify our analytical technique.

So the first thing was design. How do we make it so that consumers are going to be able to look at the mortgage elements and be able to make a more informed decision? That is, rather than focusing in on a single aspect of the loan, try to take into consideration multiple factors. And then on the basis of that, be able to make an informed decision for their own circumstance.

So in this design, we had three main parts. Part one was the loan terms. And in that, we were able to give them the basic loan terms. Then, tell them more specific loan terms. Then, tell them whether or not something could change, because consumers were very interested in knowing could their amount change or mostly could their interest change. And if the answer was yes, it can change, provide more detail about what that change was going to be.

In part two, we wanted to introduce again on this first page, the idea of affordability over a long term. What are your payments going to be once you get into this house? Are those payments going to change, and how are they going to change and what direction? Do you have escrowed accounts, or is that something you're going to have to pay outside? And then also, what your estimated taxes and insurance and other fees might be like, HOA fees.

So again, trying to lay that out in a way that one, they could see it. And secondly, that it might be things that they hadn't considered previously.

Finally, we did affordability over the short term. You might be able to make those payments. But if you don't have the money for the down payment or you don't have the money for the closing costs, that loan is actually not going to be within your realm of possibility. So we wanted that to also be very clear.

Now, we had a three stage testing phase, since we were doing two disclosures. First, the loan estimate, which replaced the TILA and the HUD-1. I'm sorry, the TILA and the Good Faith Estimate. And then the closing disclosures, which replace the TILA and the HUD-1.

So in step one, we were really focused on the loan estimate. And step two, we focus first on the closing disclosure design, and then did they work together? Because we wanted people to be able to see what the changes were from the loan estimate to the closing disclosure.

And then in step three, there were a number of other-- we did a Spanish translation of these. There were things that came out of the quant study that we had to go back and test. And there were other kinds of loans that we also wanted to test.

Now, I want to do this really quickly. It was complicated the way we set this up. We started with relatively easy loans. We wanted to introduce a task, because we wanted them to do something as opposed to just talk to us about what do they think about this loan estimate. We did a think-aloud. We followed that with a comparison. And then we also did detailed follow-up questions to verify comprehension of specific terms, and could they find things on their own.

We mixed consumers up. We had people of low literacy, of high literacy, of young age, of old age, experience, inexperience, everything, every variation we could come up with, we tried to represent in our sample population. And we included industry, since those of you who work in disclosure know that often, industry goes, wait, we don't like this. It's a change. So we wanted to make sure that we had some idea.

We also did a variety of loans. This is probably one of the most complicated comparisons they had to do. You will see that just about everything has a hedge on it. Specifically if you look at this one, you can see that the interest rate may be lower on one, it's higher on the other, but it changes at a different rate, and it changes at a different point in time. And where it can go can be higher and lower.

We made these as complicated as we could as we came near the end of the project, again, trying to test the strength of the ability of people to follow it.

Now, this is our analytical techniques. We use Bloom's Taxonomy. These are qualitative interviews that we're doing. And we wanted to make sure that we were able to have a way of articulating what was going on with consumers as they talked to us about the trade-off they were making. And we could very well say, oh, they did well. But we wanted to have something that we could put against our opinions on things.

And Bloom's Taxonomy, which some of you, or probably all are familiar with, but I understand maybe it's not the most perfect instrument. But it was a useful one for us, because in the beginning, all we wanted to know was recall. And when we tested our early stages, that's what we were seeing. People were dominant down there. A few people, of course, could go up and give us a fairly sophisticated trade-off, but most people were focused down on the bottom.

As we went through the development process, what we saw in a scatter plot, you would see more people up there in the evaluation and the synthesis stage. And that was what was important to me, because we wanted to hear the rationale around what they were deciding and how they were making the decision.

As I am running out of time, as I expected, let me go through these pretty quickly. You'll be able to see them in the slides when you get them. But we have quotes from knowledge where they just know, oh yeah, I know what an adjustable rate is, and they can tell us all the way through-- and we're not going to look at any of these-- to the point where you can see someone actually reasoning.

I could put 10% down, and that would make my closing cost a little bit higher, but I'd have a different kind of an interest rate. But maybe I should just be putting that money into my pocket, because then I have immediate access to it. Oh, but wait. Then, I won't. I'll be financing a larger amount of money, and maybe that's not the smartest thing to do. That to us was a fairly sophisticated way of trading off what was going on in this particular loan.

This one again is seeing the same type of thing. The trade-off between a closing cost and an interest rate, and taking into consideration many other factors, even if this quote is focused on one. But again, that sense that there was a richness to the thinking as opposed to what we sometimes saw when we began our study, which I will admit was shortly after the financial crisis, when people would go, oh, an adjustable rate? I don't want one of those. Those are bad. This was a much more sophisticated thinking about it.

Now, our quant study, again, relatively complicated. Done against the current disclosures at that point. And these proposed disclosures again showed a lot of richness and a complicated study in which they had to be able to compare, talk to us about why they were comparing this. So there was a qualitative question on this as well as a number of other questions about content.

We had relatively good success. And in a way, we were fortunate in that what we had was an opportunity to have a quant study that could verify the same results that we were getting as we thought about Bloom's Taxonomy and were looking and analyzing the results we were getting in these interviews.

Again, almost everything was statistically significantly better on the proposed disclosures over the current disclosures with one exception, and that was a question that we later changed. Again, not going to spend a lot of time, but you can see there is an improvement in all of these.

And I think my final takeaway is that comprehension is much more than being able to identify a word or find something in a disclosure. It is really about being able to integrate that information and be able to apply it to yourself so that you understand not merely the technical meaning of something, but the implied meaning of something. And for us, that definition of comprehension is far more important.

You can get these reports here. They look something like this. These are the executive summaries, but the full reports are there. Thank you.

[APPLAUSE]

RYAN MEHM: Thank you. Thank you, Susan. Next, we will hear from Joel Reidenberg.

JOEL REIDENBERG: Thank you. I'm delighted to have the opportunity to participate in this session today. And I really applaud the FTC for convening this workshop to explore how we can learn whether disclosures actually achieve their goals.

The work that I'm going to be talking about is co-authored work. It's part of a broader project called the Usable Privacy Project. It's an NSF-funded project with Carnegie Mellon, Fordham and Stanford where we're trying to develop automated tools to better enable users to understand privacy policies. And one of the challenges we've been finding in the work is the ambiguity in those policies. It's a real challenge to creating technical tools.

So we started working on this problem. And we have two papers that are going to be coming out. One is law and policy paper. The other is a technical paper. And I'll be speaking about the methods and what we did in each of these projects.

We had three research goals for these projects. The first was to develop some kind of theory and way of measuring and comparing vague and ambiguous terms in a privacy policy. We wanted to do that so we could test whether regulation could improve the clarity with respect to what users actually understood the policy to state.

And then we wanted to test how vagueness might affect user's perception of risk and their willingness to share personal information. We started by trying to come up with a typology for ambiguity. What did we really mean?

And we did this-- this was really the first step. We had to identify the categories of terms that we thought were going to be vague. So we came up in our research with four different kinds of vagueness that we see typically expressed in privacy policies-- conditional terms, generalization terms, modal terms and numeric quantifiers.

To give an example, this actually is almost verbatim from a real statement in a privacy policy. I think we may have changed one or two words in it. And we've annotated it. We've marked it to show what we mean by this. "We generally"-- well, generally is a generalization. "We may"-- that's a modal term. "Certain service providers, some of whom"-- those are each these vague quantifiers. "As necessary" is a conditional term.

We want to see ultimately how the use of these terms used in relation to data practices conveys information to users. What do users understand from this? So here, for example, "we generally may share." "Share" is the information practice. "Generally" and "may" are vague terms being associated with it. And we want to understand, again, what users actually learn from this. Do they understand what the website is going to do with their data?

We had the categories. Our next step was to figure out the taxonomy. What are the vague terms that we're going to look for in privacy policies to test against user reactions?

We used a method called grounded analysis, where we took a set of privacy policies. In this case, we wanted to look at employment, telecom and shopping sector. And through grounded analysis, we came up with a taxonomy of terms. So here's the illustration we show.

Conditional terms that are commonly found in privacy policies or words like "depending, necessary, appropriate, inappropriate, as needed." We again set up this whole grid of terms to look for. We then annotated a set of policies marking these terms so that we could do an analysis and survey work with users.

We had a working hypothesis that we thought as vague terms would be combined, we wondered or thought that this might increase the lack of understanding, lack of comprehension on the user's part. So we hypothesized that as you progress and add more vague terms, the senses are going to become more and more meaningless to consumers.

We thought at some point, there may be a saturation level reached, so that it would be utterly no meaning whatsoever for a user. A user would not glean from the sentence anything about the website's privacy practice.

So we wanted to test this. We used a technique called a paired comparison study. So what this does is it looks at the relative impact of vagueness on each category in combination of terms that are being used. So here's an example of the survey question. We would say, for each of these questions, which one represents a clearer, a more clear description of the company's treatment of your personal information?

And by giving a series of sentences, we can vary, and in different combinations, test each of the categories, the four categories that we looked at, and each of the actual terms, the vague terms used in the policy.

When you do this, you can then come up with a statistical measure that's known as a Bradley-Terry coefficient. And the Bradley-Terry coefficient shows the relative impact of a term compared to others. So what we could do with this is determine-- you'll see on this chart. I'm not going to go into what our results were, but more illustrate what this shows us.

The initial CNMG, those represent a conditional term, a numeric quantifier, or a generalization term. What this chart shows is when you put these terms in combinations in the same sentence associated with the data practice-- so, "We may generally share information," we have the grouping of them. When you group them together, depending on the categories you use, it either makes them more vague, which is the top right-hand corner, or less vague, which is the bottom corner.

So you can see, for example, if you take a modal term and a numeric term, you add a generalization to it, it makes it substantially vaguer than adding a conditional term. That moves it to the left. So by using this paired comparison analysis and Bradley-Terry coefficients, it gives us a way of measuring the impact on users of the use of particular terms, whether it's clearer or less comprehensible to them.

It does not tell us anything about the completeness of a policy, because it relies on the existence of terms in the policy. We used another measure to address completeness, whether a policy addressed different kinds of practices. We can also take a look using the Bradley-Terry model for terms within a category.

So here, we see a bunch of terms that are in the modal category, from "likely" to "possible." And what this tells us is we can compare them and see "possibly" will always be perceived by users as a less clear statement of the data practice than "likely." So it gives us a scaling relative to each other.

To then take a look at the clarity of a policy for users and to compare one policy against another, which is really what we were trying to do. Can we compare the clarity of policy a against the clarity of policy b? We came up with a scoring method which created a score. It's not an absolute score. It is a comparative relative score only.

And what it does is it takes essentially the sums of the Bradley-Terry coefficients for every sentence describing a data practice, and it normalizes it and gives us this vagueness measure. That enables us to compare policies against each other.

The second research question we had was comparing regulatory models. Do different regulatory models result in clearer privacy policies as understood by users? We tested it against the model privacy form that the regulators had approved. We tested it against EU Safe Harbor companies.

And then, let me turn to our third research question, which is now that we have mechanisms of measuring and thinking about vagueness, we wanted to know whether the use of vague terms combined with perceptions of risk increased or decreased the willingness of users to share information.

So we used a technique, factorial vignette survey, with users where we would look at two independent variables. So one independent variable's a vague term as measured by our earlier work.

And the second independent variable was the likelihood that somebody in close proximity to the user-- close proximity could be in their family circle, their work circle-- and spatial proximity, physically in their community, in their state, whether these two independent variables gave us some indication on willingness to share information.

So in doing the factorial vignettes, we would pose a scenario. We would replace different risk levels, and we would match it against different vagueness questions. The vague statements were coming from the earlier research showing those that were more or less vague according to our prior survey research.

And that gave us a mechanism of determining certain kinds of vague terms. We can measure certain kinds of vague terms will increase the willingness or decrease the willingness to share personal information with a site.

And I'll move to just a couple very quick ending points. We think this has a lot of application for improving clarity in statements. We can use some of these survey results for linguistic guidelines to determine what kinds of words to actually use, and in what proximity they can be used in drafting privacy statements.

We think that there are broader applications for this kind of method to consumer contracts, end user license agreements. We don't know whether our particular results with the specific terms we focused on is domain specific, or whether if you use, "we may share," whether that's going to have the same impact in other domains. But that's something that's easily tested.

And all of our papers from the project, including these two that this presentation is based on is at usableprivacy.org website. Thank you.

RYAN MEHM: Thanks, Joel. Thank you.

[APPLAUSE]

Thank you so much to all our panelists. Now, I'm going to ask some questions. So my first one, and this is open to the entire panel, is from your experience, what are some of the most effective ways to evaluate disclosure comprehension? Betsy, do you want to take that one?

ELIZABETH HOWLETT: Yeah, I'll jump in. What we find is actual understanding is if we ask them to choose the most healthful product, for example, and they're able to do that with a high degree of accuracy, I think those kind of accuracy measures are very important. It shows that they understand and want to carry out.

RYAN MEHM: Great. Dan, did you have something you wanted to add?

DANIEL GOLDSTEIN: Yeah, I wanted to say that I think, and listening to some of the other talks today, that behavioral research has gone through this sample size revolution. So my favorite way of assessing comprehension is through these very large crowdsourced studies.

And I think that participants used to be expensive, and it led to some things like if you'd get a diverse set of people, it would be expensive, so you'd ask very, very long surveys to people. And a lot of people were complaining about getting nonsensical responses back, and that's going to happen when you're asking a lot of each person.

Or you get a few people together in person, and you try to understand how groups differ from each other. And then, I think the risk is you overfit individual differences. You think the people are more different than they actually are.

But now, participants are cheap. You can go in, you can ask very few questions of each person, get quality answers back. And you can properly model these differences between people. So I think that we might make better progress by just asking more people to evaluate things because of these revolutions in subject recruitment. And I think that's what I'm really excited about, I'm going to continue with.

SUSAN KLEIMANN: If I could add, I think one of the considerations is going to be the type of disclosure, which we were talking about at the break. That there are some disclosures where there is a right answer. There are other disclosures like mortgages, and even privacy notices to

some extent, in which the answer is really dependent upon the individual, characteristics, and what their experiences were and all of the other things that we know.

And I think in those cases, it is important to really try to get a sense of what the rationale is behind the decision so we're not just dividing people into categories of, you got it right and you got it wrong. Because oftentimes, there's a much finer nuance going on in the thinking that I think it is incumbent upon us to try to capture.

RYAN MEHM: Joel?

JOEL REIDENBERG: Yeah, I wanted to just jump. I guess I second the sentiment that it's really in surveying users that becomes critical here. It's also very hard though to define the survey questions in a way that will give you the kinds of answers you're looking to find.

We did a project, an analysis of privacy policies, where we had three different groups, including privacy experts read sections of a policy and then answer questions on it. And one of the striking findings that we had is that nobody agreed on what the terms meant in the policy. That some of it was, I think in part, our study design. But some of it also was just in the difficulty in assessing the language in the documents themselves.

So there are inherent difficulties in framing your problem and framing your survey to get results that will really be helpful for you in improving comprehension.

RYAN MEHM: I want to follow up on a point that Dan made. Dan, you said-- and I'm paraphrasing here-- a few minutes ago that it's becoming cheaper and cheaper to recruit participants. Yet we also heard earlier on the last panel that an eye tracking machine costs around \$30,000. One of the other panelists I think mentioned a study that's \$5,000 that it costs.

So my question to anyone here on the panel is what recommendations or suggestions can you make to small or medium-sized firms that would like to test for disclosure comprehension? Their heart's in the right place, but they just fear that they lack the resources or they actually do lack the resources. What can they do?

DANIEL GOLDSTEIN: So I can make a couple of practical suggestions. There's a proxy for eye tracking, which is mouse tracking that we've used successfully. And you can get open source code to do mouse tracking. So this is basically to look at where people are positioning the mouse on the screen. And it is a good proxy of where people are looking on the screen, because people read with their mouse more than they think.

Another is just keep watching the frontiers of what's coming out because now, eye trackers are getting super cheap. There are things that can just basically work through a webcam in a laptop, which is pointing at you all the time anyway. People can give permission to turn that on, and it can get a good read on where the eyes are. So I think these prices are really going down if you just stay on the forefront of technology.

RYAN MEHM: Betsy did you want to add something?

ELIZABETH HOWLETT: Yes. I also wanted to add that within our discipline in marketing, we used to have to contract with some of the marketing research services. And typically, it was anywhere between \$5 and \$6 a person with Amazon Mechanical Turk. That's really simplified the process. Some of the journals were getting a little bit of pushback from some of the journals on that.

But in general, you can go out, and like Dan was saying, you can just have a few items, pay respondents a quarter. And the project's literally completed in an afternoon. Apparently, there's a whole community of these people out there.

And there have been a number of studies that actually compared the accuracy of these where you're paying \$5 or \$6 per person to the Amazon Turk, where you're paying for a reasonable sized, maybe \$0.75. And my personal experience is they're equal quality as well. So it's a very cost-efficient way. And the data that you're getting is quite good.

DANIEL GOLDSTEIN: I'll second that. There's a lot of reviews that have shown the quality is equal, and sometimes better on Turk.

ELIZABETH HOWLETT: And I will say the one thing that we didn't realize when we started using it the first time is now, we have to specifically state, do it on a laptop. Because you saw the stimuli. It's kind of complicated to show a picture of a cereal box or something. We noticed that there are a lot of people taking it on cell phones.

And so I would just alert people that really, if you're doing something more complicated, to make sure that they're taking it on a laptop or on something not a cell phone. But if cell phone is good, that's good for them.

RYAN MEHM: Joel.

JOEL REIDENBERG: I was going to suggest that small and medium-sized businesses could partner with university research centers. Because in many instances, they don't need to reinvent the wheel. Some of the basic questions they have are probably already answered by some research findings that they're not going to be aware of. And issues that they have that really should have some new testing could be of great interest to university researchers as well.

And the university researchers are much more accustomed, I think, to doing the Mechanical Turk surveys than a typical small, medium-sized business person is going to be. So there will be mutual interests between businesses and university research that could be advantageous to both.

SUSAN KLEIMANN: And from a qualitative perspective, we work with a lot of large government agencies, but also small businesses. I guess my key message would be it is better to test anyway than to not test.

And so one of the things that we will sometimes suggest is if you must, at least ask some employees to sit down and read through things. You want to be able to think about it. It's not going to be a publishable study. But at least you're going to get some feedback on where the

trigger words are or where a letter informing people that, oh, you've been exposed to radiation testing, and we didn't tell you at the time. How are you going to convey that kind of information?

I think the other thing that we have done under very tight time frames is going to a Starbucks and give people a card. Now, the caution there is that you want the demographic profile of the Starbucks you are going to match the people that you actually want to reach. Because otherwise, you can be way off. Choosing a Starbucks on Capitol Hill is probably not going to give you a population of lower income people.

RYAN MEHM: So let me ask you, Susan. How do you recruit participants for disclosure comprehension studies that are representative of the types of consumers for whom a disclosure is the most relevant?

SUSAN KLEIMANN: So we do, again, it's qualitative research, and we want to be there in person. And we are often being asked to get geographic diversity. So we do use marketing firms. It is not inexpensive. We work with our client to develop a profile, a demographic profile of age, and ethnicity, race, gender, income, whatever those factors are that they want us to be considering, and then we try to get a mix of that.

Usually in a large study, as the CFPB study was, what we want is at the end of the study, we actually have a demographic profile that is representative. But at the beginning, you're not necessarily going to be able to get that at each individual site, especially given that you're rarely doing more than 10 interviews. Pretty hard to get 10 people to represent a national profile.

RYAN MEHM: Thanks. I want to turn now to icons, and that's something that we've heard about throughout the course of the day. Ilana talked about earlier this morning. Betsy touched on it in her remarks. Other panelists have talked about icons as well.

And when disclosures are presented as simple icons, such as the Healthy Selection icon tested in Betsy's product packaging study, there may be a risk that consumers misunderstand the meaning of the icon or read into it things that don't hold. So Betsy, my question for you is did you explore what people's thoughts and conceptions were of the Healthy Choice or Healthy Selection icon? And then I'm going to turn this to the rest of the panel with some questions for them about icons.

ELIZABETH HOWLETT: That's a great point. And that's one of the shortcomings when you have an evaluative icon, such as this is healthy.

There's what Brian Wansink has declared as a health halo. And so you have this halo that's surrounding the product. And so for example, maybe it's low in sodium, but people are also going to assume that it's low in fat. They're going to make these sort of assumptions. And that can be a real issue. And yes, we did test for it. And yes, it does exist.

RYAN MEHM: So I want to ask the rest of the panelists if they have experience testing comprehension of other types of icons. And if so, what did you find?

JOEL REIDENBERG: We did a project that was doing some preliminary testing for a plugin that would show information to users about the privacy policy, what the privacy policy says. And I think our results were generally confused, is about the best way to say it. Because when you start showing icons, one of the difficulties-- and we had a lot of just in designing the study, was a extensive process with computer scientists and our legal team.

There are so many different elements that you have to make judgment calls as to what goes into the icon that we don't really have a public consensus around them. So when you present it to a user, there isn't a natural understanding that a user would have as to what x, y or z might mean.

So if you think about if you want to give a letter grade, for example, to a privacy policy for its content, you need to have a consensus on what are the criteria that go into the letter grade. What's the relative weighting of different factors in that criteria? So if it's opt-in, how do you weight opt-in compared to the individual having the ability to correct wrong information? You need to have some agreement on how to weight them.

And those factors make it very complicated to-- we just don't have consensus to show a letter grade to show some other score, it tends to be very confusing. And in our testing, we didn't have anything that clearly emerged as an obvious choice.

RYAN MEHM: Thanks. Joel, you found in your study, you clearly found that there are specific words that are particularly vague and confusing for consumers. So my question to the rest of the panel is have any of you studied comprehension of individual words? And if, so could you talk about that?

ELIZABETH HOWLETT: Well, one of the things that we looked at within the context of nutrition is one thing that really started us on this whole comprehension interest, is that we were doing some initial work when the FDA was considering adding trans fat to the nutrition facts panel.

And we did some research, and we had for whatever reason, we had the sodium and trans fat on the same front of package. Oh, three grams of trans fat. Oh, that's low. That's good. It's like, oops. No, that's not good. And sort of just that there's no reference point. So consumers really don't have a reference point, which is what we talked about. How many calories does it take to work off 50 calories or 250 calories? There's no reference point. And so without that, you need that.

RYAN MEHM: Well, thank you. We are just about out of time. So I'm now going to convey to you the most important disclosure of today's workshop, which is we will take a lunch break until 2:00 PM. Lunch is available for purchase in the cafeteria. Note that as Lorrie mentioned this morning, if you do leave the building for lunch, you will have to go through security again when you return.

We would like to encourage you to join FTC staff and speakers in the cafeteria for discussion over lunch. And note that food and drink may not be brought back here into the auditorium. And I know there's a lot of options down the hall, a lot more than cups of shredded lettuce.

[LAUGHTER]

So thanks so much.

[APPLAUSE]

[MUSIC PLAYING]