How to Conduct Eyetracking Studies

Kara Pernice and Jakob Nielsen

WWW.NNGROUP.COM | 48105 WARM SPRINGS BLVD., FREMONT CA 94539–7498 USA

This report is a gift for our loyal audience of usability enthusiasts. Thank you for your support over the years. We hope this information will aid your efforts to improve user experiences for everyone.

This report is free, but it is still copyrighted information that may be updated from time to time, so please don't distribute this file or host it elsewhere.

Do not link directly to the PDF file (which may be rehosted at different cloud services). Instead, we encourage you to distribute the following link to this report's summary page on our website, in order to reduce PDF shock and to allow people to decide whether to download it themselves:

www.nngroup.com/reports/how-to-conduct-eyetracking-studies

The research for this report was done in 2009, but the majority of the advice may still be applicable today, because people and principles of good design change much more slowly than computer technology does.


Authors: Kara Pernice and Jakob Nielsen

Contents

Executive Summary

In traditional usability testing without eyetracking there are many cues our test users give us that help us evaluate designs, such as the time a person spends on a page, the action he takes there, what he reads aloud, what he hovers the mouse over, whether he smiles or grimaces, and what he comments on. These user behaviors comprise, in our opinion, the most important aspects of learning about usability. But eyetracking (ET) adds another dimension. Watching what people are looking at adds a level of interest to usability studies. Following the eye at work makes you feel as though you are in the user's head, thinking with him. If nothing else, the ET studies are more interesting and hold observers' attention more easily.

Eyetracking technology also helps test facilitators avoid a major blunder: interrupting the user when he is quiet for a long time. This is a huge benefit since interrupting the participant too much or inappropriately is one of the greatest challenges for even seasoned usability professionals.

In addition to helping to avoid a major facilitation gaffe, eyetracking technology enables us to understand more than clicks and pauses. We learn finer points about what draws attention and why. And by studying what people look at and do not look at, we are able to glean more detailed insights about how the Web user works his or her way through usable and unusable designs. Behaviors such as *exhaustive review* (where people look repeatedly at areas that seem as though they'd be helpful but are not) and *selective disregard* (where people purposefully tune out areas of the website at given times) are very apparent when watching the user's eye at work. We see which interface items are *miscues* (erroneously calling attention) and which areas trick users and are thus dropped like a *hot potato.* (All of these and more behaviors are described in our book titled *Eyetracking Web Usability*, Nielsen and Pernice, New Riders Publishing 2009.)

Conducting valid usability research calls for good planning and experience. But conducting valid usability eyetracking research is on another plane, adding to the preparation, facilitation, and analysis. Eyetracking technology takes a regular usability test and makes it more difficult, more time-consuming, and more expensive. Conducting or analyzing results of a research study erroneously is worse than not doing any research at all. Unfortunately, ET technology makes it very easy to get misleading results where you believe something that simply isn't true about your design.

In this report we discuss what we have learned about how to plan, conduct, and analyze an eyetracking study that will produce sound results that you can trust and use to improve your website. While we see many benefits to eyetracking research, we also know that good usability studies do not call for high-tech solutions. Some of the best, iterative tests are done with paper and pencil only, because the faster you can run a study the sooner you can adjust the direction of your design project.

METHODOLOGY, NOT FINDINGS

This report does not present any of the *findings* from our ET projects—it's purely about *methodology*: giving advice for how you should run your own ET studies, should you decide to do so.

More information about the findings from our eyetracking research:

Eyetracking Web Usability, a book by Jakob Nielsen and Kara Pernice

How Users Read on the Web: The Eyetracking Evidence, a report by Kara Pernice and Jakob Nielsen

SHOULD YOU USE EYETRACKING IN YOUR USABILITY STUDIES?

Probably not. The average company does so little usability that it is better served by sticking to simpler (and much cheaper) usability methods, such as thinking aloud and paper prototyping:
www.nngroup.com/reports/paper-prototyping-training-video

Only after you've conducted about a hundred rounds of regular usability testing do you reach the stage of insight where you need to dig even deeper and pay closer attention to those details that are only revealed through eyetracking.

All the big things—like people getting lost on your site, not understanding the content, or not even understanding your value proposition in the first place—can be found without eyetracking, and that's where you should focus first, if you're only going to run a couple of rounds of user research.

Luckily, some companies do have a full-fledged usability process in place and conduct large numbers of usability studies. For such companies, eyetracking can become a valuable component of the usability toolkit, if it's used correctly. Sadly, most eyetracking studies are done wrong, which is why we decided to publish this report to help you gain more value from eyetracking by employing valid methodology.

(See page 153 for a discussion of why we do eyetracking studies.)

Eyetracking Method Tips

Each tip is described in the pages that follow. This list includes only the tips themselves.

Recruiting Users for Eyetracking Studies

Recruiting the test users for an eyetracking (ET) usability session is much like recruiting for any usability session. We won't cover all the basic recruiting points here. Instead, refer to our report, *How to Recruit Participants for Usability Studies*.

But there are some additional factors to consider and things you must do that are specific to eyetracking.

COMMUNICATE TO THE PARTICIPANT DURING THE SCREENING INTERVIEW

In addition to all the typical points you communicate to any potential usability tester in the screening (such as, "We will ask you to use the Web and try to do different things, observers will be watching you work," "We will be videotaping the session," "This is in no way a test of you or your abilities—rather you are helping us test the designs,") you should also disclose to the user that his eye will be tracked during the session. You tell him this for a few reasons, including:

- Ethically, you should disclose to anyone participating in an experiment what is being done to them, especially something somewhat out of the ordinary.
- It prepares the user so when he arrives at the lab there are no surprises, and you can comfortably get on with the session.
- People are smart, and they will wonder why you are asking so many questions about their eyes in the subsequent screening questions. If you don't explain why, they may feel something fishy is going on.

But how to talk about the eyetracking to potential users is a bit touchy. Over-explaining the eyetracking can be detrimental in two major ways. First, it can make users too aware of their eye movements. If they are thinking too much about the fact that people will know what they are looking at, then they might, when they come to the lab, try not to look at things that could be embarrassing or personal. (By the way, if the users could truly train themselves this way in our study, we probably never would have seen so many of them looking at body parts, blood, and shamelessly gawking at a scantily-clad Lindsay Lohan!)

The second danger of over-explaining eyetracking during a screening is that it could simply freak the potential testers out so much that they don't want to participate at all.

You certainly do not want to make a big deal out of the technology you'll be employing. We once observed an ET facilitator who thought it would be a good idea to call it "this really cool thing that I can use to capture your eye," and then went on to overly-thoroughly explain the system. She was attempting to do something in between trying to make it sort of a fun topic and trying to explain the intricacies of the system before the session started. Her reasoning was good, to basically make sure the users understood they were being tracked and help them to not be afraid of the technology. But users really don't need to know the intricacies of the system. In this case, making the technology the focus backfired. It was like when a little kid falls and scratches his knee. He gets up unscathed mentally, but when one of his buddies points out, "You're bleeding!" the kid starts to bawl. There's no need to point out that someone is bleeding—leave well enough alone.

Out of the ten users we watched in her lab, at least four of them had some major bumps related to their curiosity or nervousness about the eyetracker. In fact, one woman just couldn't get past the fact that her eye was being tracked. She kept saying things like "That blinking is hurting my eyes," or "Is that infrared going to hurt me?" She ended up stopping and leaving after half-heartedly attempting a few tasks. It does not have to be this way in eyetracking sessions though. In fact, in the sessions we ran, nobody had any of these issues, and all were fully aware that we were tracking their eyes. Recruiting well was the first step in this success. (This topic is further discussed in the "Beginning the Session and Calibrating the User's Eyes" section on page 72 of this report.)

1. **When screening a participant over the phone before he comes to the lab, always disclose that you will be using technology that enables you to track his eye during the session, so you can see what he is looking at on the Web.**

2. **Be careful not to highlight the eyetracking this topic too much, as it can have adverse effects.**

HOW TO SAY IT

You should briefly mention the eyetracking technology just before you begin asking the questions specifically about the eye in the screener questionnaire. (See the "Screener Questionnaire" section below for more information.) That should be the potential participant's first introduction to the idea of the eyetracker. For example:

"In this study we will also be observing what you look at as you use the Web. So we will use technology to track your eye movements. Now I'll ask you a few questions related to the eye."

After all of the questions are asked and you have determined that the user has the right background for your study, you accept him into the study and then succinctly say a few more words about eyetracking. See the example below:

ACCEPT:

Your profile fits our needs for this study and I would like to invite you to participate. As mentioned earlier, we will compensate you a sum of $100 at the time of the study. The session will last for about an hour and-a-half. At the session, we will ask you to work on the Web, and we will ask you questions about the experience as you are working.

One thing that is very important for you to know is that this study is about the design of websites and not in any way a test of you or your ability. So there is no need to be anxious or to practice using the Web or computer before you come.

The session will be observed by one to three people. It will be videotaped, including mouse movements. We will also be tracking your eye movements as you use the Web, so we can see what you are looking at as you work.

Do you have any questions about any of this?

SCHEDULE AN APPOINTMENT.

The available dates right now are between Mon October ___ and Tue October ___, 10am-6pm. If you can tell me approximately what time you would like to come, I could give you the available slots to choose from.

Wait for response; schedule an appointment.

GIVE LOCATION OF THE TEST AND DIRECTIONS.

You will need to bring a picture ID to show at the building's security desk.

You participation is very important to us, and we are relying on you to make your scheduled time. Please arrive ten minutes prior to the session start time to fill out consent forms and pre-session questionnaires.

I will send you an e-mail with these details. Would you please confirm the address with me? In case of an emergency, please call Mary Keohane as soon as possible at 212-555-0000.

WHAT TO ASK THE USER

We prefer to ask the most basic screening questions first, and if the user gets through those, we move on to the eye questions. You could certainly start off with the eye questions and follow with the others, but we feel it is better to ease people into these, as we it could be odd to just start asking probing questions about their eye health.

After asking some of the most basic screening questions to determine whether the user is part of your target audience, then ask questions about the eye. First, tell them that the upcoming questions will be about the eye, and briefly tell them why you are asking about the eye at all.

3. **In the screener, ask the general target profile questions first, and if the user is a match then ask questions related to the eye.**

4. **Never in any screening questions disclose to the potential participant the answer you are hoping for before you ask the question about the eye (or any other question).**

5. **Limit the number of questions you ask about the eye to eight or fewer.**

Never in any screening questions disclose what you need or the answer you are looking for before you ask the potential participant the questions. For example, do not say, “We cannot include people in our study if they wear bifocals. Do you wear bifocals?”

The reasons you should not do this are probably obvious, but here they are: some users aim to please and will say, sometimes subconsciously, anything you want to hear. And unfortunately there is a small number of people who just want to come and get the honorarium and will tell untruths about their background to do so.

The following is a good way to introduce and ask the questions.

Screener Questionnaire

Note that the prompts to the right of the responses, such as *CONTINUE* and *Skip to 3* are for the recruiter's information only and should not be read to the user.

Eye Information: I am now going to ask you some different questions. These are about your eye because we are going to use some simple technology to track eye and mouse movements during the study.

1. Do you wear contacts or eyeglasses in order to read the computer screen?

[ ] Yes CONTINUE

[ ] No Skip to 3

2. Are your glasses for:

[ ] Reading only CONTINUE

[ ] Seeing distant objects only CONTINUE

[ ] Both (Do you wear bifocals, trifocals, layered lenses, or regression lenses) TERMINATE

3. Can you read a computer screen and the Web without difficulty with your contacts and/or eyeglasses on?

[ ] Yes CONTINUE

[ ] No TERMINATE

4. Do you have cataracts?

[ ] Yes TERMINATE

[ ] No CONTINUE

5. Do you have any eye implants?

[ ] Yes TERMINATE

[ ] No CONTINUE

6. Do you have Glaucoma?

[ ] Yes TERMINATE

[ ] No CONTINUE

7. Do you use a screen reader, screen magnifier or other assistive technology to use the computer and the Web?

[ ] Yes TERMINATE

[ ] No CONTINUE

8. Are either of your pupils permanently dilated?

[ ] Yes TERMINATE

[ ] No CONTINUE

You can ask more questions, but we found the above amount to be about the limit when also considering all the other background questions we had to ask.

It would also be valuable to know if the person wears false eyelashes, a lot of mascara, funky thick-rimmed glasses, or refuses to remove his lucky Red Sox hat. But we found these questions to be too intrusive, not common occurrences, and just plain too odd to ask when recruiting for a usability test of corporate websites. You may decide otherwise depending on your target users.

TELEPHONE IS EASIER THAN E-MAIL WHEN RECRUITING FOR EYETRACKING STUDIES

6. Recruit over the phone, not email, for eyetracking studies.

For some studies we can do recruiting by e-mail. This can often be faster on our end and from the potential participant's perspective. E-mail allows people to review the test information and screening questions at their leisure. Of course, if people are not sold on the idea of testing, e-mails are easier to ignore. But if you have run an advertisement and you have many respondents, sending them an e-mail with screener questions can work quite well.

That said, we much prefer phone over e-mail for eyetracking study recruiting, mainly because the technology is different and users may have more questions about these sessions. Also, there are some more questions we have, mainly about the user's eye, that could need clarification. That is easier done in a conversation rather than in e-mail back-and-forth.

Besides the actual recruiting element, sending potential test participants information over e-mail before the study allows them to spend some time with the information. If you have a particularly long or confusing consent form for people to sign, consider e-mailing it to them before the study so you don't waste valuable time in the lab. (Really, try to get your legal department to boil down the document to a few simple paragraphs. It's been done!)

Particularly for new usability testers and people who cannot easily read a long form—like people with low literacy, people who are blind, or people who have low vision[1]—sending them any forms in e-mail before they come to the lab can save a lot of time. And the users won't feel that we are springing some big legal document on them when they arrive on-site, and only giving them a minute with it when they require more time. This is particularly important when testing with children[2] where the parents need to sign the forms and can be protective, bordering on paranoid, because there are so many unethical studies masquerading as market research out there.

---

[1] For more on how to conduct usability sessions with users with disabilities, please see our report, How to Conduct Usability Studies for Accessibility.

[2] For more on our studies with small kids, see Children (Ages 3-12) on the Web. For teenage studies, see Teenagers (Ages 13-17) on the Web.

How Many Users to Include in a Study

THE NUMBER OF TEST PARTICIPANTS NEEDED IS DICTATED BY DESIRED OUTPUT

Before deciding the number of users to include in an eyetracking usability study, first consider what the main deliverable for your study will be. The output that has the most impact on the decision about the number of users is heatmaps. Heatmaps are fast to create, sexy, and interesting. But heatmaps can be dangerous because they appear to be qualitative representations of multiple users' fixations, when in reality they are quantitative because they are based on statistics. If you are using heatmaps to actually draw conclusions based on an aggregate of users' experiences, or if heatmaps are the main deliverable, then eyetracking requires many more test users than traditional usability studies. If using heatmaps to analyze data, ensure that you have 30 users per heatmap. Thus, you should include about 39 users (as opposed to five or so for a traditional qualitative study).

The number of users needed makes heatmaps the most expensive of the common user research methods, as shown in the following table. Even quantitative measurement studies without eyetracking are substantially cheaper, though admittedly the estimate of needing 20 users is for each design you want to measure. Usually, you benchmark in order to make a comparison, and with two designs, you'd need 40 users, making such a study marginally more expensive than eyetracking, if you only wanted heatmaps for one design.

Research Method	Users	Reference
Qualitative user testing (thinking aloud)	5	*Why You Only Need to Test With 5 Users*
Card sorting	15	*Card Sorting: How Many Users to Test*
Quantitative user testing (measurement benchmark)	20	*Quantitative Studies: How Many Users to Test?*
Eyetracking aimed at generating heatmaps	39	*Eyetracking Methodology, 2009* Kara Pernice and Jakob Nielsen
Qualitative eyetracking (watching gaze replays)	6	*Eyetracking Methodology, 2009* Kara Pernice and Jakob Nielsen

The table above shows the recommended number of test participants to include in each study for several of the main user research methods. Note that the numbers for the two eyetracking methods include an allowance for users whose eyetracking recordings are of too low quality to be used because of the limitations of current

eyetracking technology. We need good eyetracking data from 30 users for a heatmap and 5 users for a qualitative review of the gaze replays, and if we hypothetically had perfect eyetracking technology then that's all we would have to test. In the real world, the numbers in the table are the current recommendation.

7. **If you want to draw conclusions using heatmaps or if heatmaps are the main deliverable for the study, you need 30 users per heatmap. This mean you'll actually need to include about 39 people in the study.**

8. **We recommend watching the sessions and the gaze replays from sessions for the best analysis, not heatmaps. In a qualitative ET study, about six users is typically enough for an accurate assessment of the main usability issues in a design.**

VARIABILITY IN HEATMAPS DEPENDING ON THE NUMBER OF USERS

The reason we need more users is that there is more variability when you try to quantify minute detail the way we do in eyetracking. Take the homepage for the U.S. Census Bureau. On this homepage, users exhibited four different classes of behaviors: search-dominant, navigation-dominant, tools-dominant, and the successful behavior of reading the population count right off the Population Clock. To draw an accurate heatmap, you need all four behaviors represented, in approximately the correct proportions, or you'll get the wrong colors on parts of the screen.

About half of our users were search-dominant on the Census homepage. Let's say that the true percentage, if we had tested millions of people, comes to exactly 50%. Now, if we only test 10 people, we would expect to have five search-dominant users in the sample. But that's not the way statistics work. If you toss a coin 10 times, it doesn't necessarily come up heads exactly five times, even if it's a fair coin.

You only get five heads from ten tosses 25% of the time. 41% of the time, you get either four or six heads, and 34% of the time you get zero through three or seven through ten heads. In other words, the "average" outcome is actually the least likely outcome. A slightly skewed count of heads is the most likely result, but even a highly skewed count is more likely than getting exactly five heads. Note that 34% of the time we expect a result that would lead people who don't understand statistics to suspect a fraudulent coin. But because a fair coin produces a weird result 1/3 of the time, you can't actually conclude anything if you get, say, eight heads out of ten tosses.

Variability doesn't matter if we're running a qualitative study, because all we would report to the client is that "many" users depend on the search and what particularly striking difficulties they had using it. That's all we need to improve the search feature. It doesn't matter whether it's 40%, 50%, or 60% of users who depend on search in this scenario—in any case, it'll cost too much lost business if we don't fix the search. Thus, the client should spend their remaining money on fixing their search, not on paying us to run hundreds more users to narrow down the confidence interval on the estimate of search-dominant users among their target audience.

For a heatmap, however, we must know how many people looked at the search box, and for how long, in order to decide how red to color that area of the screen.

We might easily color something yellow that should have been blue if we're plotting a heatmap based on a small number of test users. That's why eyetracking studies become very expensive if their entire goal is to produce heatmaps.

Let us look at three concrete examples:

- A search engine results page (SERP) from Google
- A corporate homepage from Adelphia
- A section page from *New York* magazine

For all three examples, we collected good eyetracking data from 60 users, so we can plot the heatmaps based on this large sample size. We can also divide the sample up in smaller groups of users and plot the heatmaps as if we had only tested a smaller number of people. For the sake of illustration, we will look at heatmaps based on 10 users each, meaning that we can plot six different heatmaps for each of our examples. (Because we have 60 users, we can divide them into six non-overlapping groups of 10 users, meaning that we can simulate having run six different eyetracking studies with 10 users.) We will also look at heatmaps based on 20 users and heatmaps based on 30 users. (Because we have 60 users, we can plot three different sets of 20 users and two different sets of 30 users, still without having any overlap between the groups.)

We will do all of this both for the heatmaps representing the users' first 10 seconds on the page and for the heatmaps representing the users' entire visit to the page.

First, let's consider how people look at search results on Google. During the first 10 seconds, we can see from the heatmap from the full study that users look a good deal at the first two organic results (in other words, they get almost five seconds a piece) and that there's a small bit of viewing of the first sponsored ad.

If we only based our heatmap on 10 users, we might think that there was *a lot of viewing of the ads*, since that's the case in one of those heatmaps. But if we had recruited a different set of 10 users, the conclusion might have been that there was *no viewing of the ads* at all during the initial 10 seconds, because that's what one of the other heatmaps shows. The remaining four heatmaps all show a bit of viewing of the ads, though one of them shows more viewing of the second ad than of the first. Our conclusions regarding the effectiveness of search engine advertising would have been completely different, depending on what 10 users we had tested. This again shows that we can't rely on heatmaps from only 10 users.

Regarding the organic listings, most of the heatmaps based on 10 users show that people read two listings during the first 10 seconds, but one heatmap shows people looking at many more listings. So our conclusion regarding organic listings would also not have been reliable if we had based it on a heatmap from only 10 users.

Even when we look at the heatmaps from 20 users or 30 users, we still don't get a firm picture for whether users look at the ads during the initial 10 seconds. Some heatmaps say "a lot," some heatmaps say "a bit," and some heatmaps say "no looking." Even 30 users are not enough to get a stable heatmap for the initial behavior on the search results page.

The first 10 seconds viewing a Google SERP on www.google.com. Heatmap based on 60 users.

The first 10 seconds viewing a Google SERP on www.google.com. Each heatmap is based on a different group of 10 users.

The first 10 seconds viewing a Google SERP on www.google.com. Each heatmap is based on a different group of 20 users.

The first 10 seconds viewing a Google SERP on www.google.com. Each heatmap is based on a different group of 30 users.

Now let's look at people's entire visit to the search engine results page. With all 60 users, we see the expected "F" pattern of viewing, with extensive reading of the first two organic results, a bit less reading of the next three to four organic results, and a tiny bit of reading of the last results. There's a reasonable amount of attention paid to the first two ads, and very little attention paid to the last two ads.

If we drew the heatmap based on only 10 users, our conclusions would vary widely, depending on which 10 users we happened to include in our study.

In some heatmaps, it looks like people read five results carefully. Others show a highly scattered amount of attention across an even larger number of results, and yet others show that people only read two results. The reason for this difference is obviously that there is no single number of results that everybody reads. Overall, the tendency is for users to read two results carefully *on average* followed by quicker viewing of three or four results—again on average. But some users happen to feel lucky and are satisfied by a much smaller number of results for a given query, whereas other users happen to be particularly critical about certain results sets, which they inspect with great scrutiny. If you only test 10 users before drawing your heatmap, you're susceptible to erring one way or the other, depending on whether you happen to get a few more users than the average from either of these two categories.

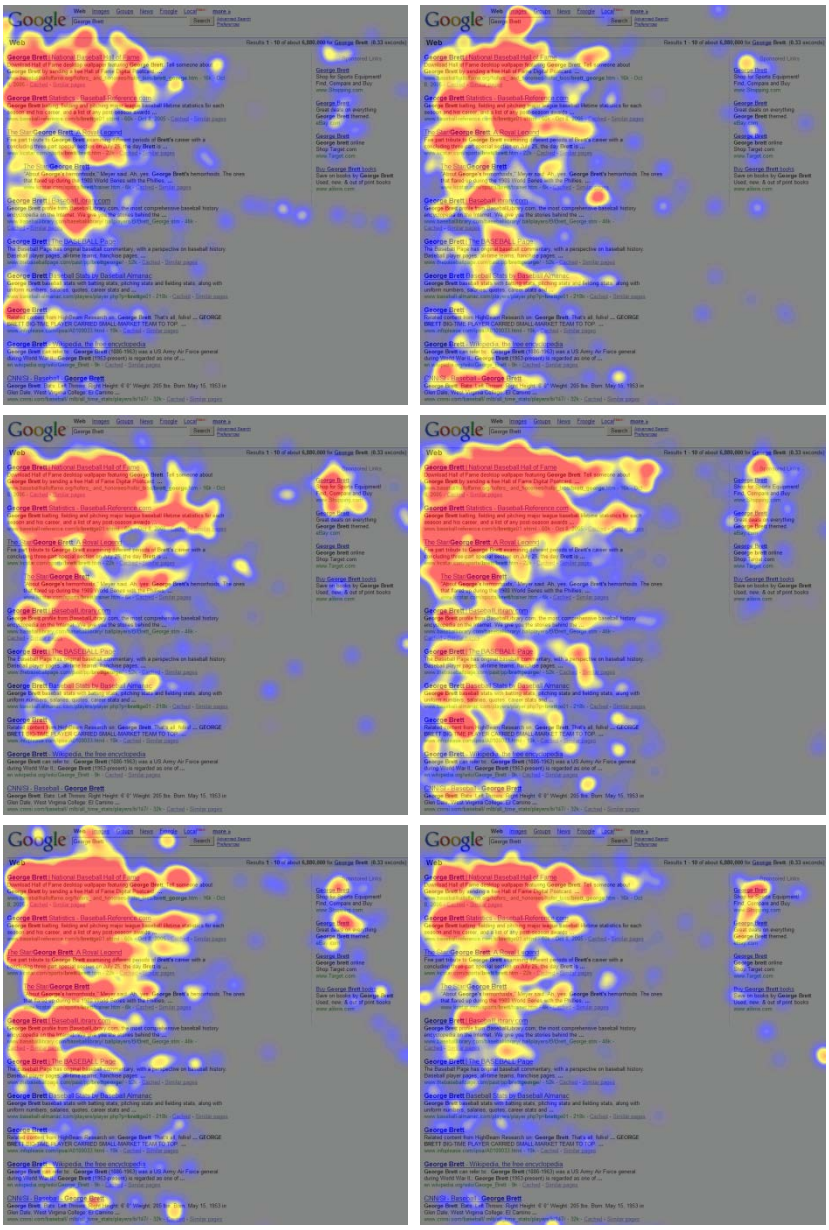
Similarly, with only 10 users, we might conclude that users hardly look at the ads at all; we might conclude that they care significantly about the first ad but not the rest; or we might conclude that they read a lot of ads. The *average* may be two, but there's great variability.

With 20 users, the shape of users' attention to the organic results has somewhat stabilized, but there's still one of the three groups that would lead us to erroneously conclude that people spend a lot of time on the last results and another group that would lead us to erroneously conclude that they almost never look there.

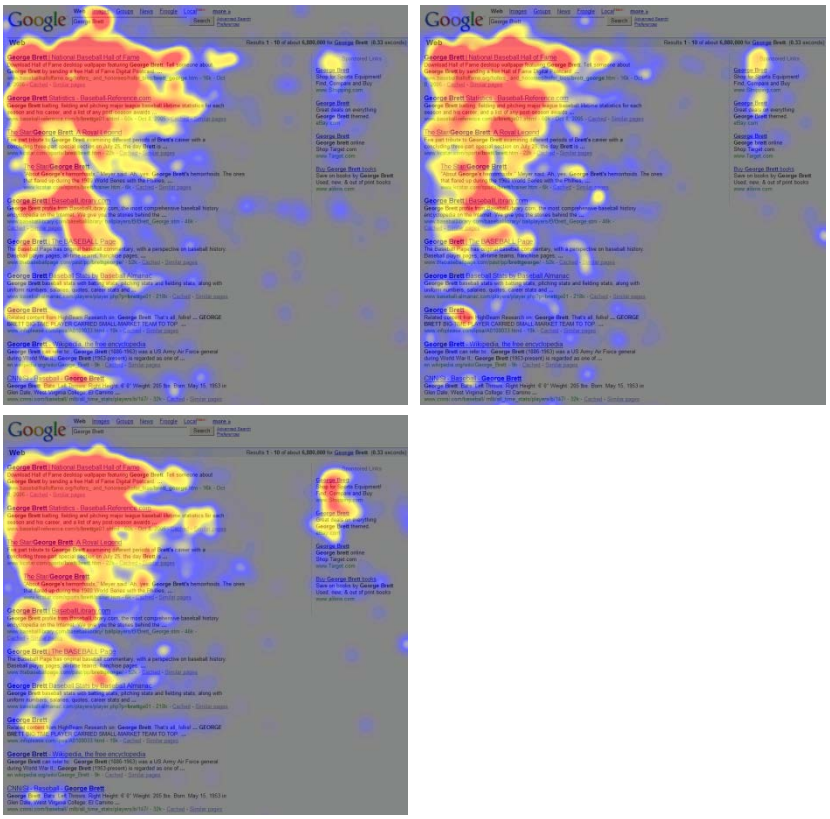
With 20 users, we no longer risk concluding that users didn't look at the ads at all, but we still don't have the amount of attention to the top ad nailed down.

We need to move to 30 users before we can say that the two heatmaps are sufficiently similar that almost all important conclusions would be the same, no matter which group of users we happened to have tested.

The entire visit to a Google SERP on www.google.com. Heatmap based on 60 users.



The entire visit to a Google SERP on www.google.com. Each heatmap is based on a different group of 10 users.



The entire visit to a Google SERP on www.google.com. Each heatmap is based on a different group of 20 users.

The entire visit to a Google SERP on www.google.com. Each heatmap is based on a different group of 30 users.

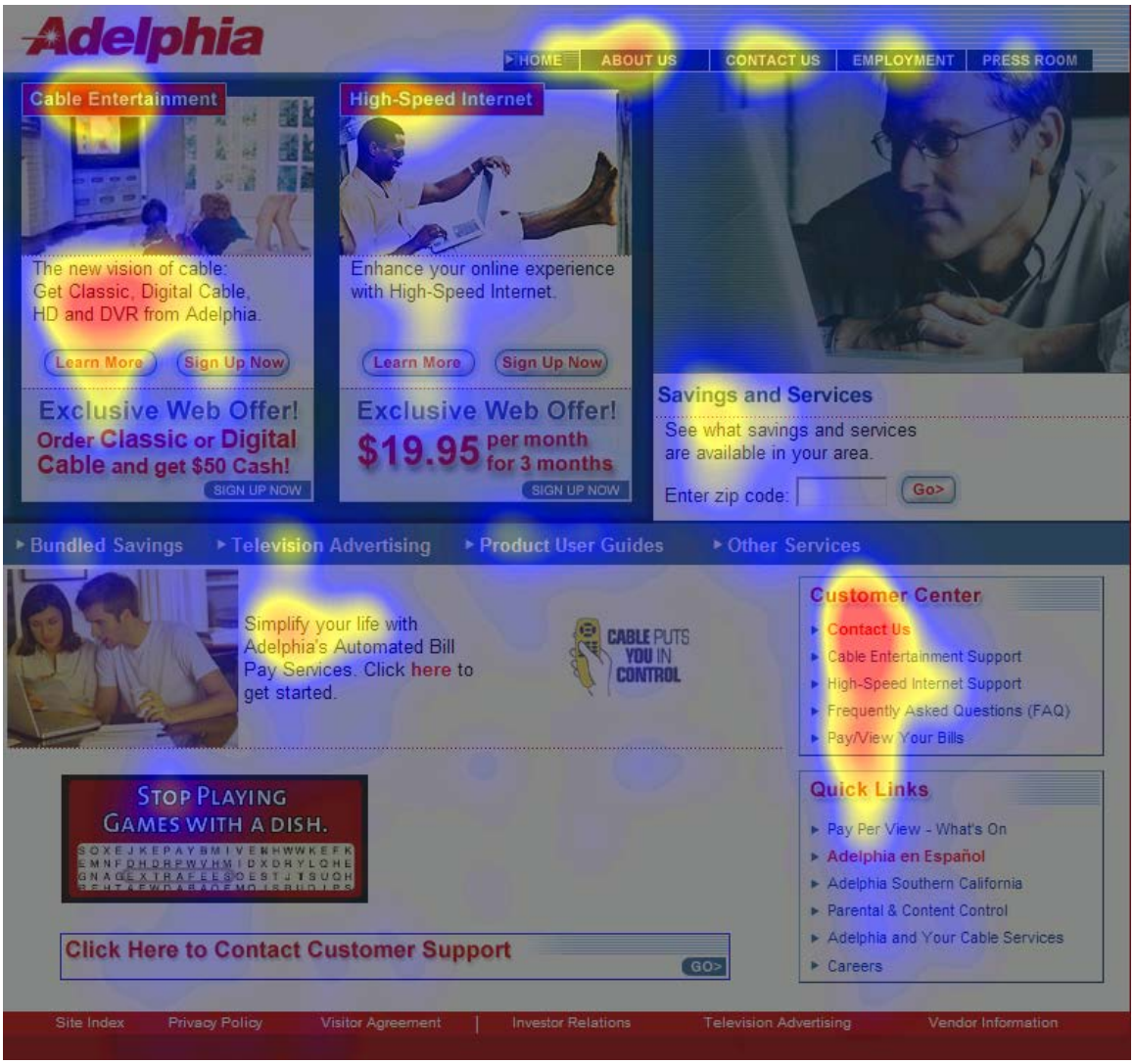
Let's look at a non-search example: the corporate homepage for Adelphia, a cable TV service. First we consider the heatmap for the first 10 seconds users are on the page, in order to analyze their initial reaction.

With all 60 users, the heatmap shows a large amount of viewing of the bulleted list in the lower right, a large amount of reading of the blurb about cable entertainment, and a small amount of viewing of the top navigation menu, the headers for cable entertainment and high-speed Internet, as well as some reading of the first 1/3 of the blurb about automated bill payment. Note that users were trying to decide whether they wanted a digital video recorder (DVR), which is a product under cable entertainment (the acronym is even included in the blurb). Thus, the analysis of the first 10 seconds shows that users are able to focus on the relevant part of the page within this short amount of time. We also note that users don't allocate any of their initial attention to any of the multiple images on the page, except for a tiny amount of fixations on the *"stop playing games"* promotion at the bottom of the page.

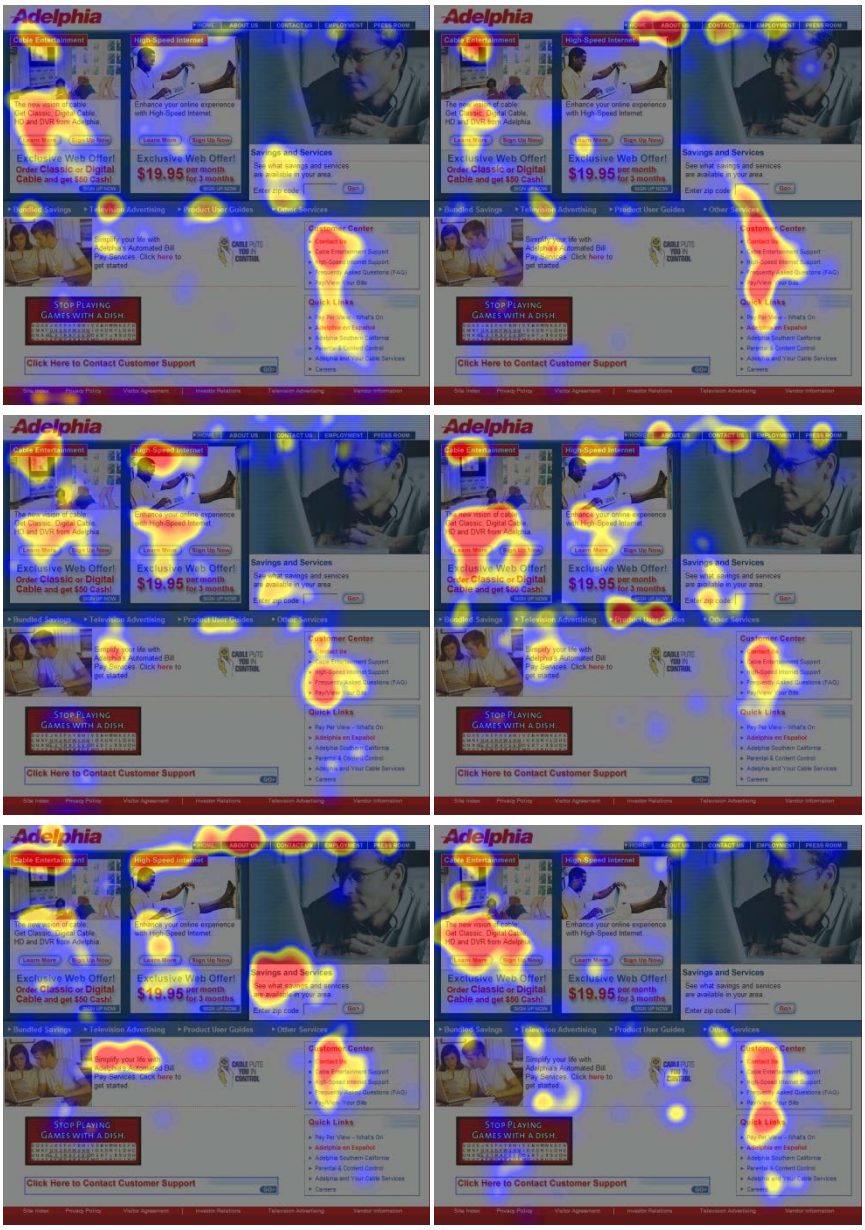
If we only have 10 users, the conclusion is not as clear. For some sets of 10 users, we see extreme attention to the cable entertainment area of the page, but for other sets, the heatmaps show almost no viewing in this area. A few heatmaps even show a tiny amount of viewing of the image of two happy laptop users, belying the general conclusion that the images basically weren't looked at. And we have one heatmap where there was more viewing of the Internet area than of the cable entertainment area of the page. Also, for the bulleted lists in the lower right, some heatmaps show that users only scan the top list, whereas other heatmaps show some scanning of the bottom list as well. In other words, with only 10 users, the heatmaps for the initial 10 seconds lead to widely different conclusions, depending on what users we happened to test.

With 20 users, the heatmaps start looking fairly similar, and most of our conclusions would be the same, no matter which 20 users we tested. Still, in one heatmap, there's virtually no viewing of the Internet area, whereas another heatmap shows a fair amount of viewing of this area.

Increasing the size of our study to 30 users doesn't help us much in this case: Do people look at the blurb under the big photo in the upper right during their first 10 seconds on the page? One heatmap says yes; the other says no. Do people look at the Internet promotion? One heatmap says yes; the other says no.



The first 10 seconds viewing the corporate homepage on www.adelphia.com. Heatmap based on 60 users.



The first 10 seconds viewing the corporate homepage on www.adelphia.com. Each heatmap is based on a different group of 10 users.

The first 10 seconds viewing the corporate homepage on www.adelphia.com. Each heatmap is based on a different group of 20 users.

The first 10 seconds viewing the corporate homepage on www.adelphia.com. Each heatmap is based on a different group of 30 users.

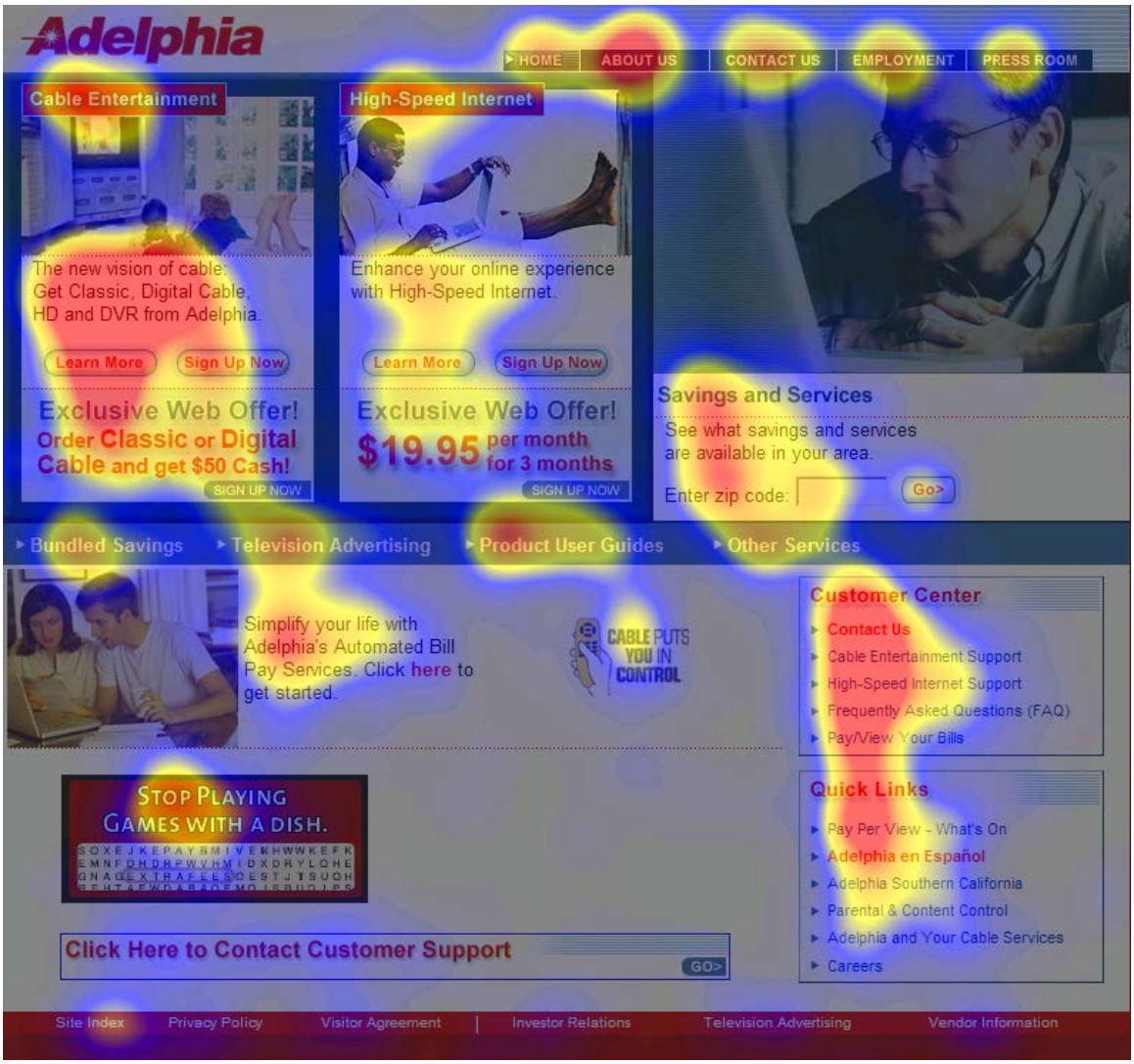
Now let's consider the heatmap for the users' entire visit to the homepage. The heatmap for all 60 users shows pretty much the same behavior as for the initial 10 seconds, just more of it:

- The users mainly read the info about cable entertainment (which leads to the product they are considering.)
- Users spend much less time on the other main product area, high-speed Internet.
- Users spend a lot of time on the bulleted lists in the lower right. And for their entire visit to the page, they even read most of the bottom list, which got very little attention during the initial 10 seconds.
- People look at the navigation menus. And for their entire visit to the page, they even look at the menu at the middle of the page. (During the initial 10 seconds, only the top navbar got looks.)
- There's some viewing of three of the promotions: automated bill payment, *"cable puts you in control,"* and *"stop playing games."* The bottommost banner for customer support doesn't get looked at. (It's probably the victim of banner blindness because of its shape.)
- There's still no viewing of the many photos.

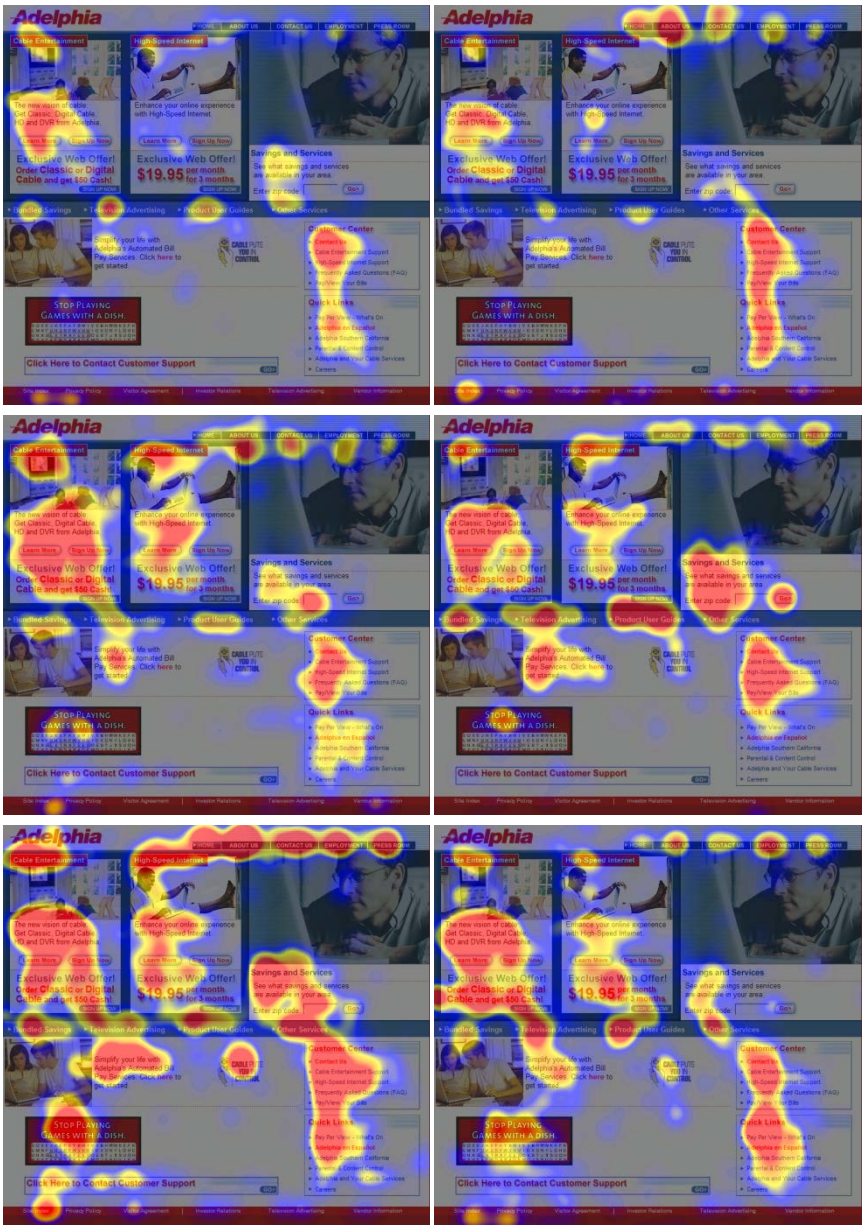
The heatmaps drawn from only 10 users don't allow us to confidently conclude that any of these findings are accurate. In some heatmaps, various promos don't get fixations. In some heatmaps, the bottommost bulleted list either doesn't get looks, or it gets very few, whereas other heatmaps show this list as being viewed just as much as the bullets above it. Some heatmaps show no viewing of the Internet area, whereas others show substantial viewing of this area. Some heatmaps even show a fair amount of viewing of the footer area, which in the 60-user heatmap shows up as getting a tiny amount of viewing.

Increasing the user base to 20 users helps and removes many of the spurious findings, but we still can't say that everything has become clear. For example, the *"stop playing games"* promo gets extensive fixations in one heatmap and almost no fixations in another heatmap. For the two bulleted lists, one of the heatmaps shows no viewing of the bottommost list.

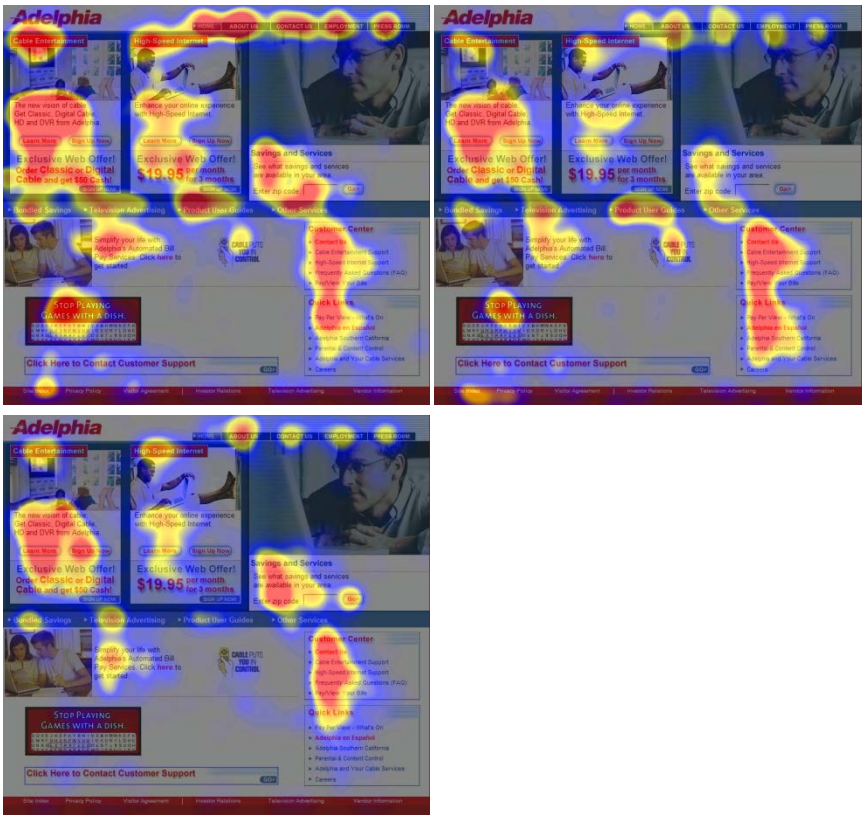
Finally, moving to 30 users, we see that the two heatmaps support the same conclusions, with one minor difference: whether users look at the *"cable puts you in control"* promo.



The entire visit viewing the corporate homepage on www.adelphia.com. Heatmap based on 60 users.



The entire visit viewing the corporate homepage on www.adelphia.com. Each heatmap is based on a different group of 10 users.



The entire visit viewing the corporate homepage on www.adelphia.com. Each heatmap is based on a different group of 20 users.

The entire visit viewing the corporate homepage on www.adelphia.com. Each heatmap is based on a different group of 30 users.

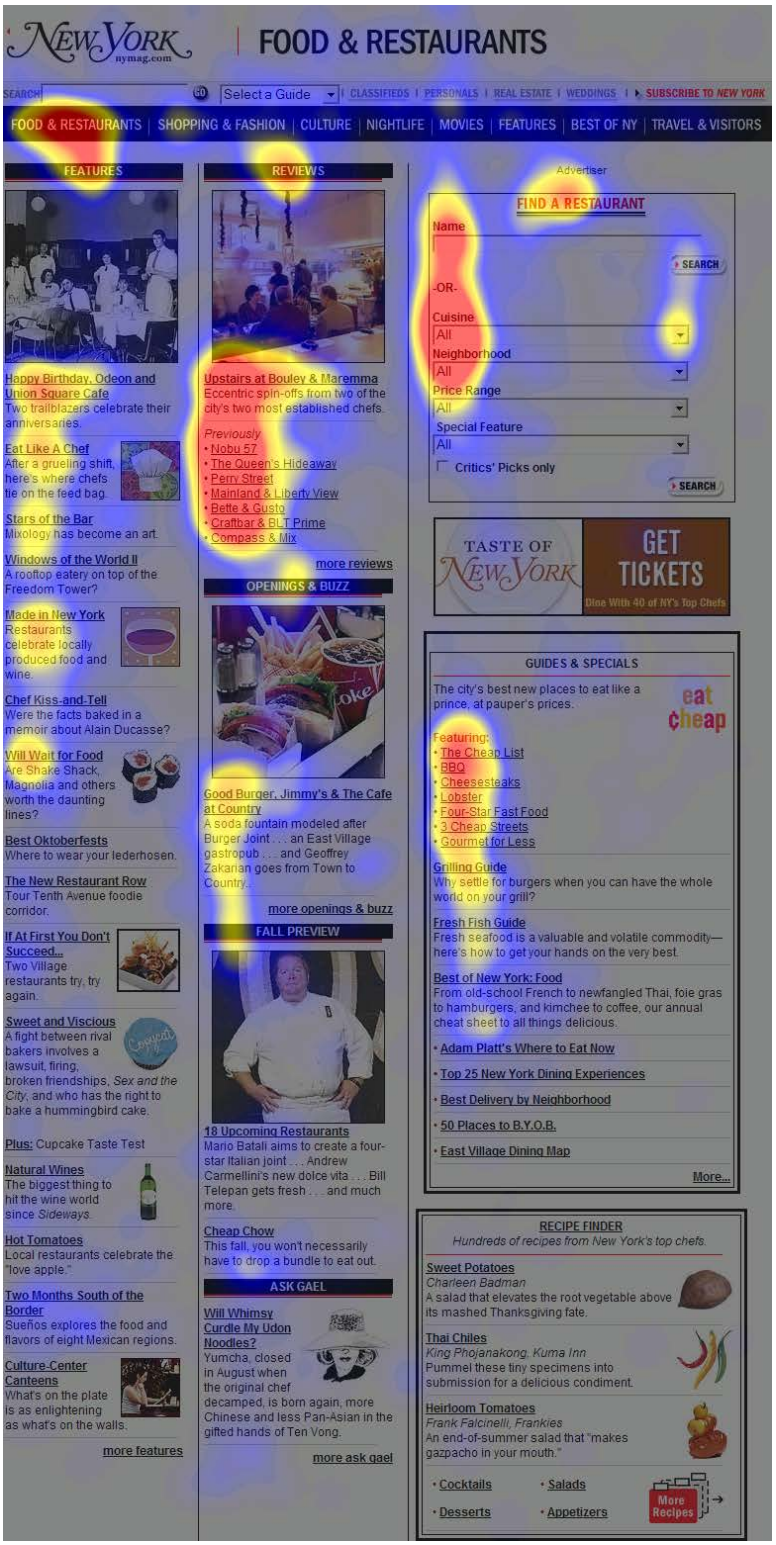
For our last example, let's look at a content site and analyze the Food & Restaurants page on *New York* magazine's site. Again, we will analyze heatmaps both for the initial user experience (the first 10 seconds on the page) and for users' entire visit to that page.

During the first 10 seconds, the heatmap for all 60 users shows a large amount of attention to the bulleted list of restaurant reviews in the second column. People also look a good deal at the *Food & Restaurants* menu, the *Find a Restaurant* tool and the bulleted list of *Guides & Specials*. There's a smaller amount of attention to the beginning of the first column and the second part of the second column. Finally, there are almost no fixations within the images or the bottom half of the page.

OK, how many of these findings can we deduce from heatmaps based on only 10 users? Very few. None of the individual heatmaps shows all of these conclusions. All of the heatmaps show attention to the *Find a Restaurant* tool and the bulleted list of restaurant reviews, but even for these strong findings, you might be unlucky and get 10 users who didn't look at these two areas much during their first 10 seconds. As for *Guides & Specials*, some groups of 10 people viewed this area a lot; other groups hardly viewed it at all. We even have two heatmaps that show substantial viewing at the bottom of the page.

Moving to heatmaps drawn from 20 users, our conclusions are much more solidified. The main trouble comes from concluding whether users look at the first column during their first 10 seconds on the page. Two heatmaps say yes; the last says no.

With 30 users behind a heatmap, the picture is very solid, and we would draw the same conclusions, no matter which group of 30 we had tested. Again, the first column causes a bit of trouble, with the first heatmap saying that it's viewed a little and the second heatmap saying that it's viewed a lot. But mainly, we'll be fine if we employed 30 users to test this page.

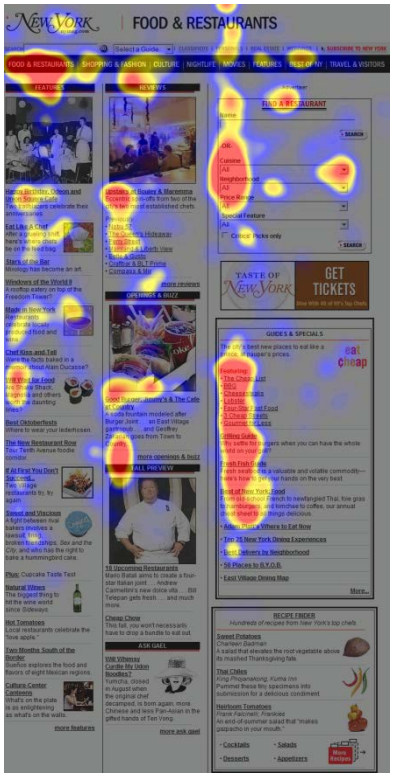


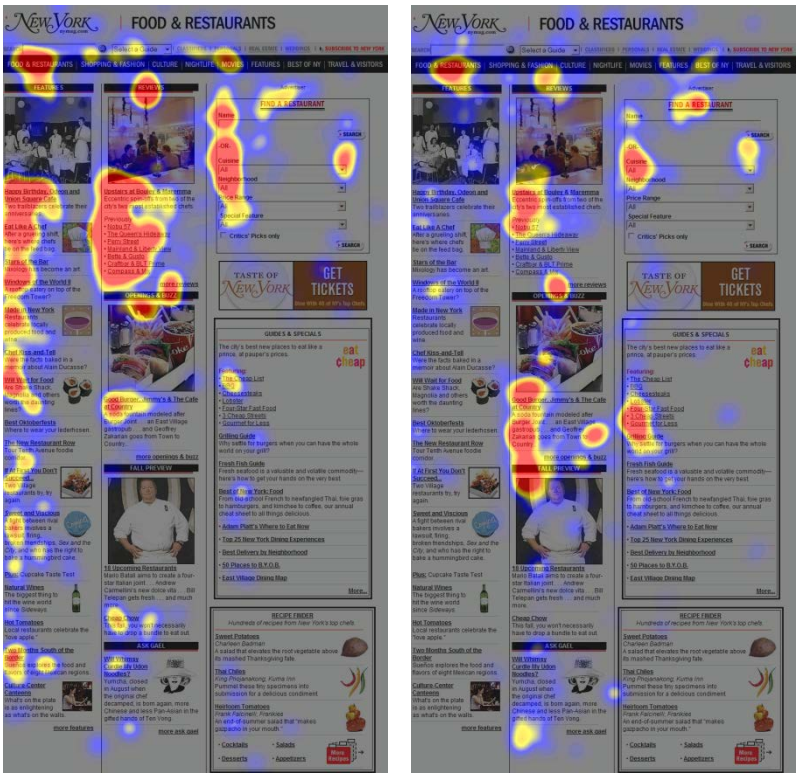
The first 10 seconds viewing a category page on *New York* magazine's site, www.nymag.com. Heatmap based on 60 users.


FOOD & RESTAURANTS

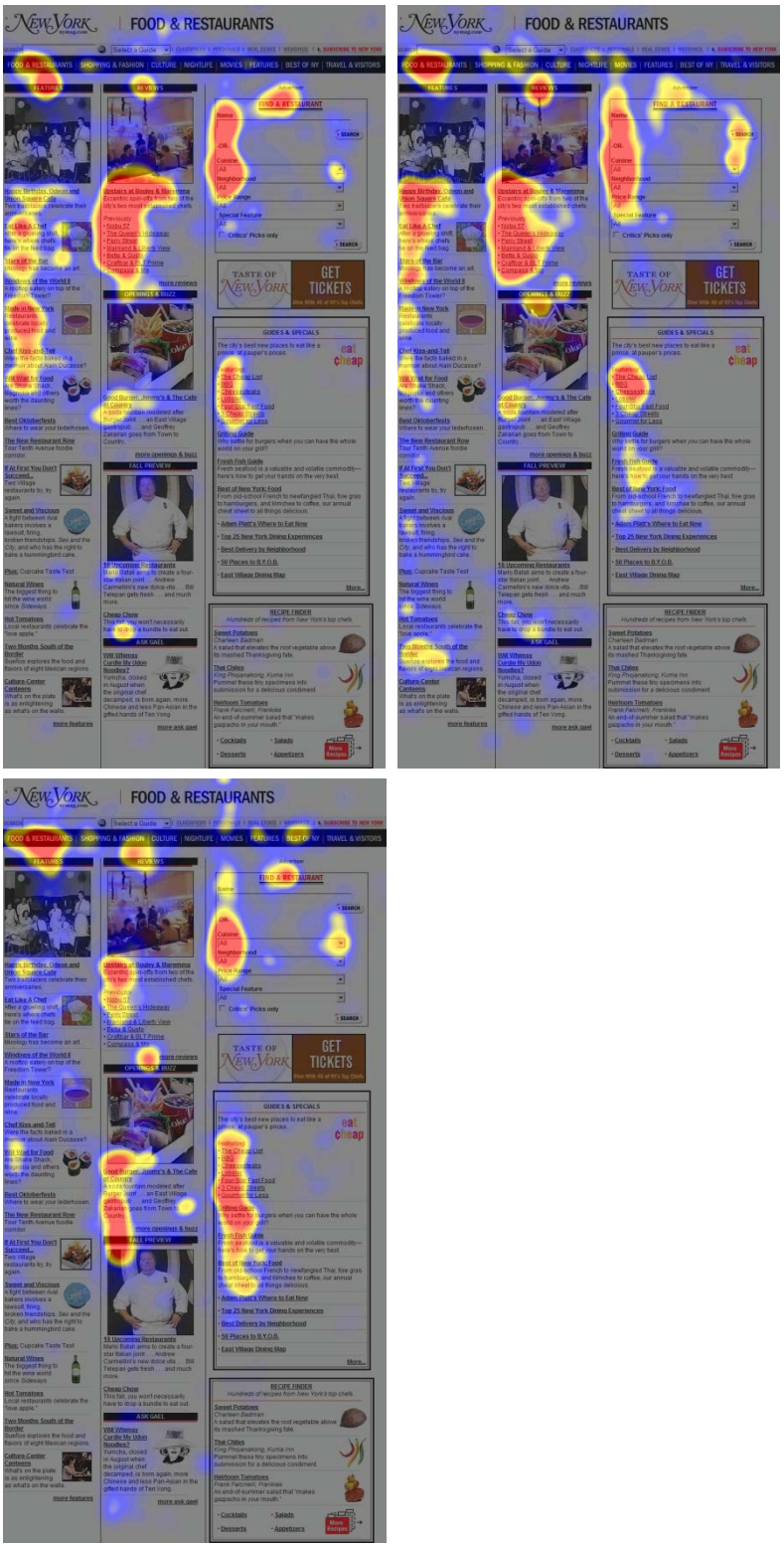

FOOD & RESTAURANTS


FOOD & RESTAURANTS

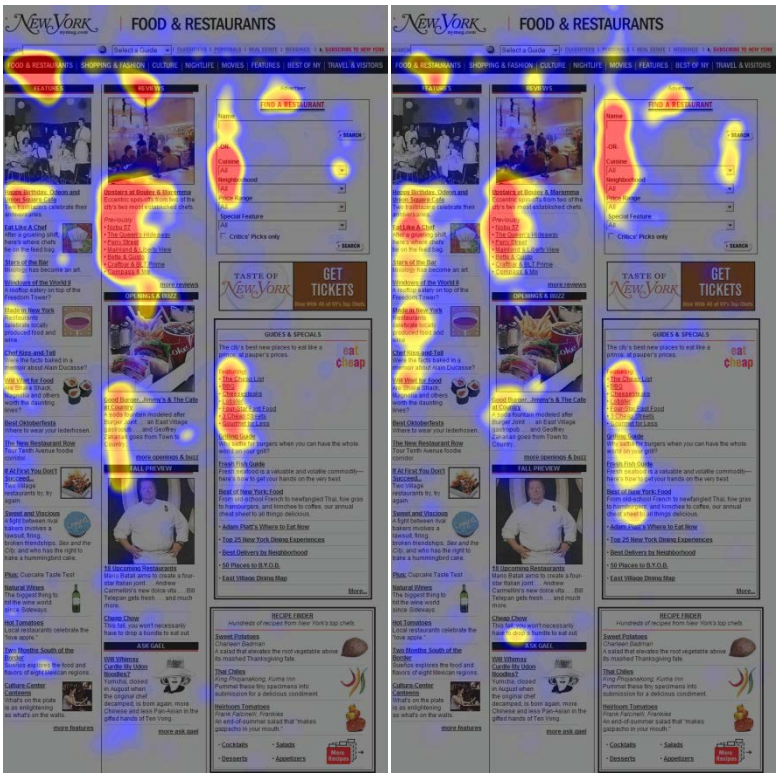

FOOD & RESTAURANTS



The first 10 seconds viewing the corporate homepage on *New York* magazine's site, www.nymag.com. Each heatmap is based on a different group of 10 users.



The first 10 seconds viewing a category page on *New York* magazine's site, www.nymag.com. Each heatmap is based on a different group of 20 users



The first 10 seconds viewing a category page on *New York* magazine's site, www.nymag.com. Each heatmap is based on a different group of 30 users.

For the users' entire visit to the *New York* page, the conclusions obviously change relative to their initial 10 seconds on the page. In fact, the conclusions change more for this content page than they did for the search page and the corporate homepage we dissected earlier in this section. Now we see strong attention to five areas of the page:

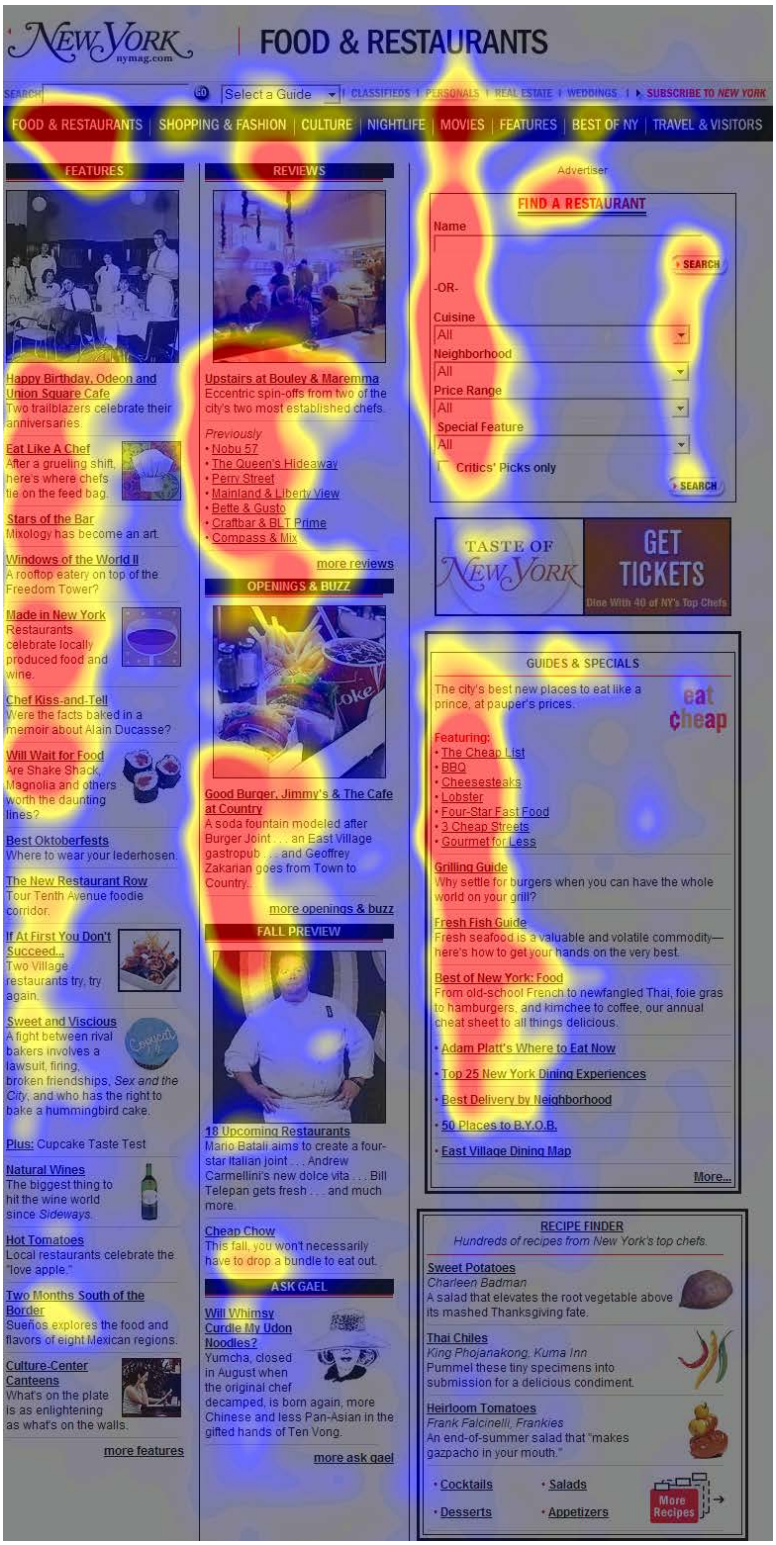
- Top half of the first column
- Bulleted list of restaurant reviews
- Second part of the second column, the *Openings & Buzz*, where users read most of the blurb, but didn't look much at the photo showing some fries (people probably know what fries look like, so this isn't a very compelling image)
- *Find a Restaurant* tool
- *Guides & Specials*, which was scanned to the end, except for the very last item

The top navigation bar is viewed a fair amount, and there is a small amount of viewing of the bottom third of the page, though this drops off quickly, with no views at the very bottom of the page. Banners, ads, and most of the images get no or very little viewing.

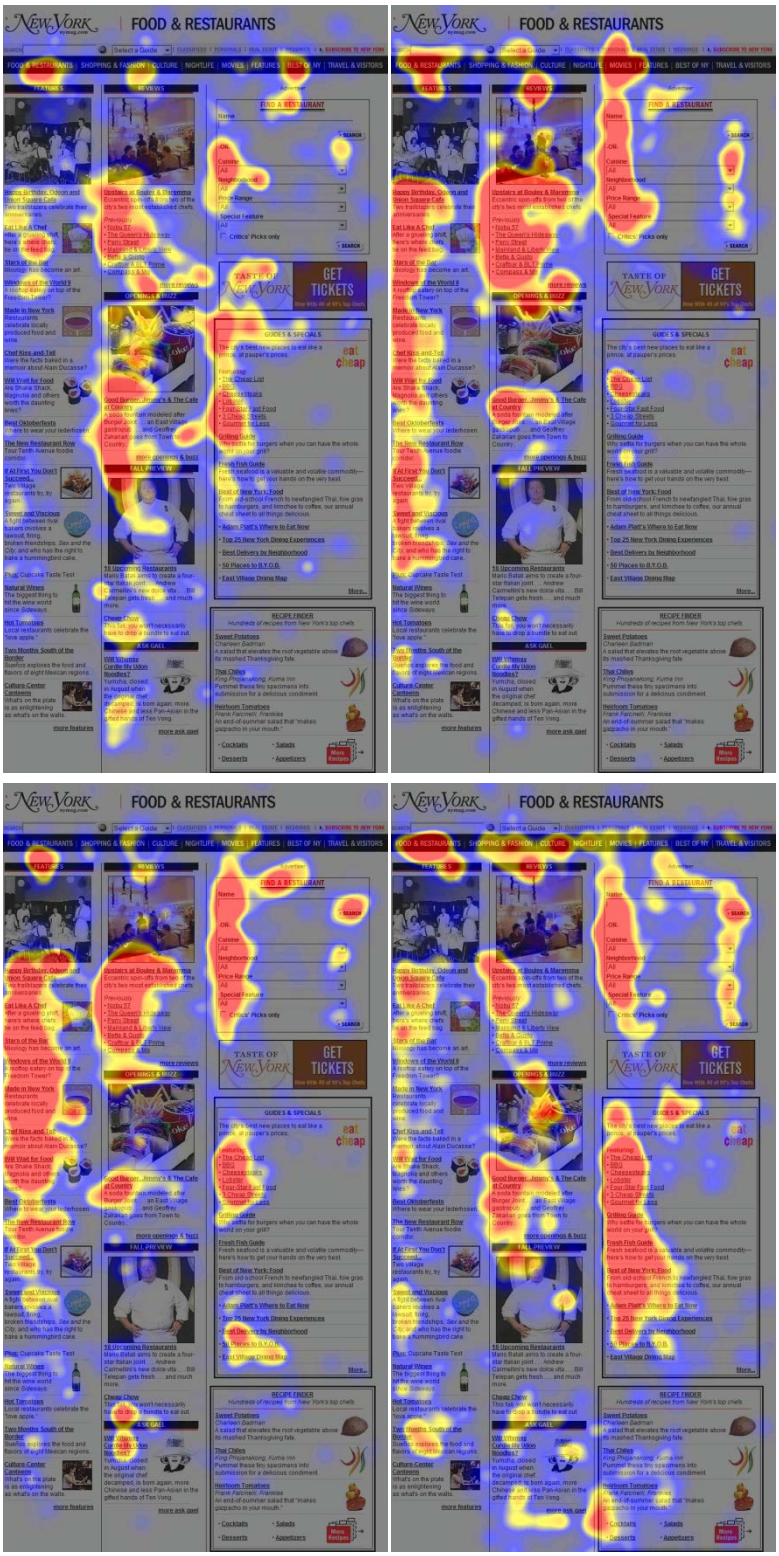
What if we only had 10 users for our heatmap? We would be in deep trouble regarding *Guides & Specials*, with half the heatmaps showing little viewing and half showing extensive viewing. The first column continues to be a source of uncertainty, with one heatmap showing almost no viewing, and two showing fairly limited viewing. There is even one heatmap that shows a decent amount of looks at the very bottom of the page. In total, in many cases we would draw the wrong conclusion if we drew a heatmap from only 10 users.

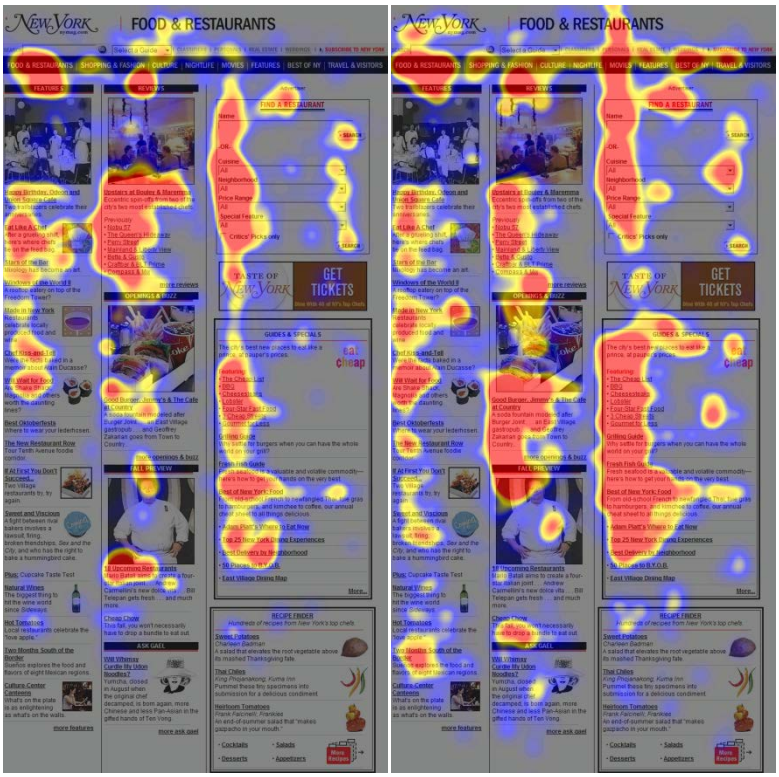
Moving to 20 users, the situation improves, and most of our conclusions would be correct. The first column continues to cause trouble, and one of the heatmaps would actually lead us somewhat astray to a belief that people scanned this column more ruthlessly than they tend to do. We also continue to get a somewhat varying amount of attention to *Guides & Specials.*

Finally, if we draw heatmaps from 30 users, the picture has stabilized so much that all of our conclusions are firm, with one minor exception: How much do people read the bottom half of the first column? One heatmap says a good deal; the other says hardly at all.

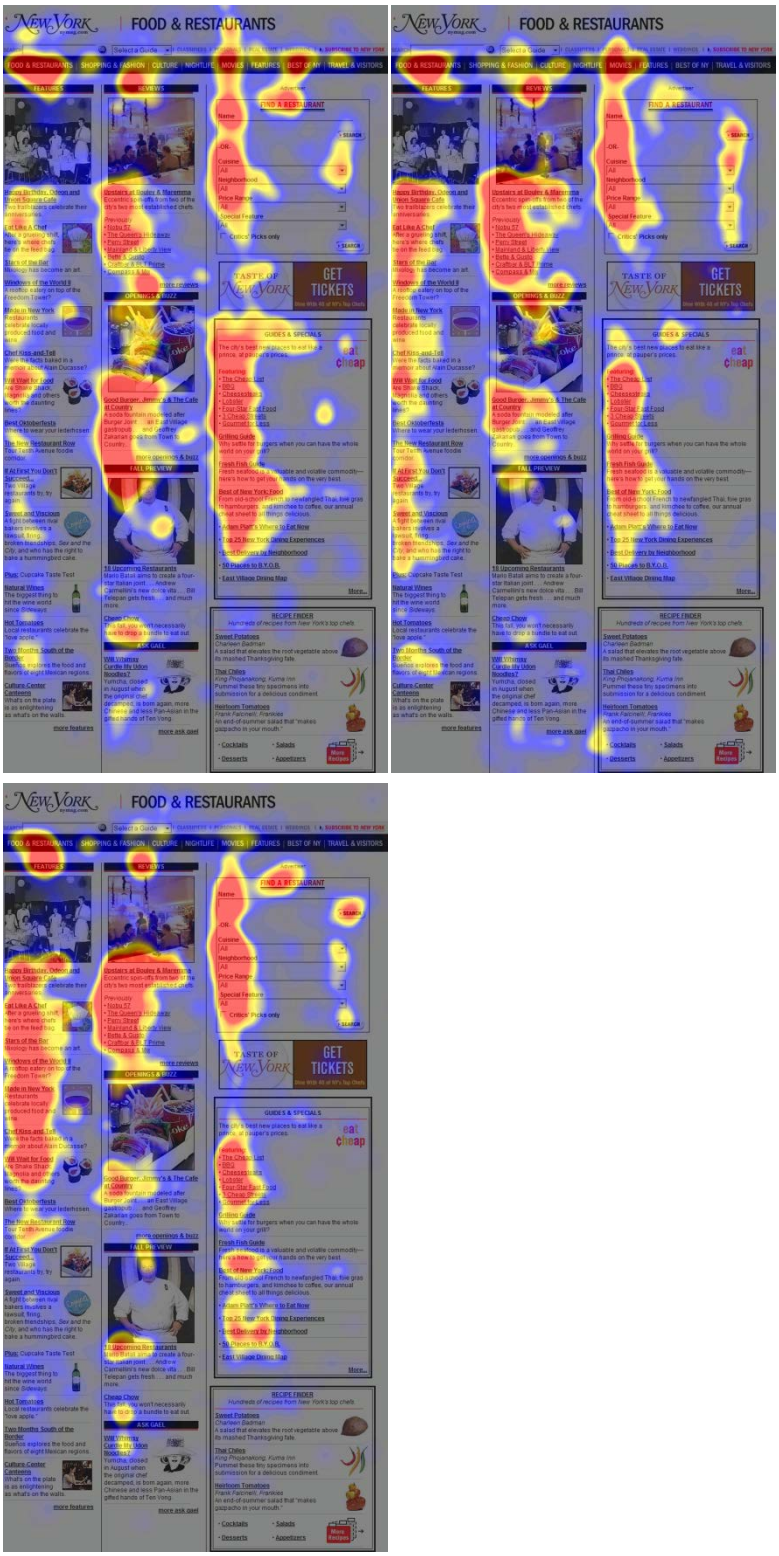


The entire visit viewing a category page on *New York* magazine's site, www.nymag.com. Heatmap based on 60 users.

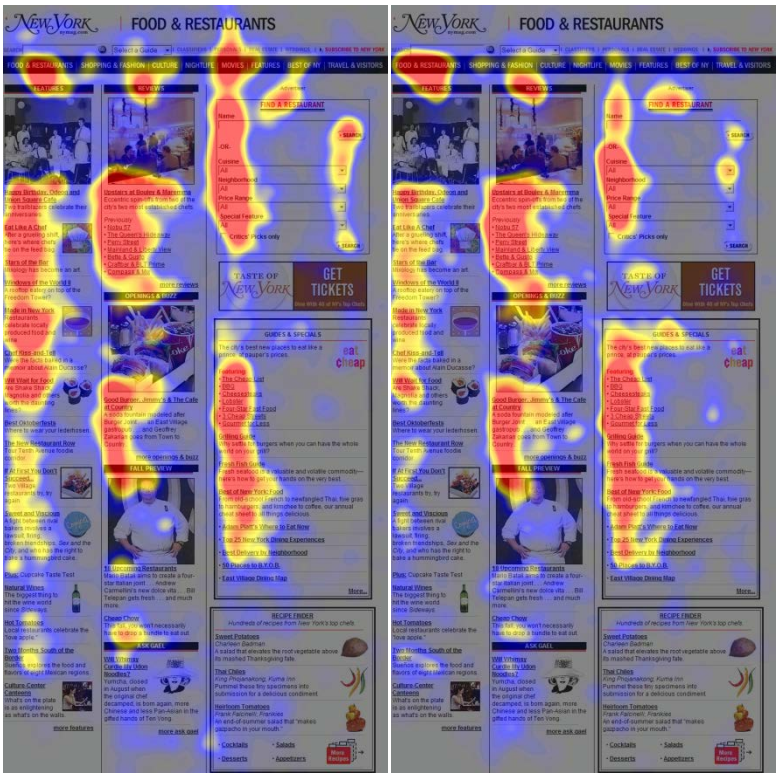

NEW YORK
FOOD & RESTAURANTS



The entire visit viewing a category page on *New York* magazine's site, www.nymag.com. Each heatmap is based on a different group of 10 users.



The entire visit viewing a category page on *New York* magazine's site, www.nymag.com. Each heatmap is based on a different group of 20 users.



The entire visit viewing a category page on *New York* magazine's site, www.nymag.com. Each heatmap is based on a different group of 30 users.

Across the three examples we have considered here, it would often be enough to have 10 users for a heatmap if we were only interested in the very broadest of conclusions. But there is too much variability in users' viewing behavior for the heatmaps to have stabilized sufficiently for this to be recommended. And any kind of detailed conclusions are certainly not safe if they're made solely from a heatmap from 10 users.

As we would expect, 20 users get us better data, but we have to move to 30 users before the heatmaps really start to stabilize.

THE SLOWLY-DIMINISHING RETURNS CURVE

The previous section looked at three different examples of Web pages and used a qualitative analysis to consider how safe we could feel drawing conclusions from heatmaps based on various numbers of users. Now, let's quantify this question.

Instead of asking how much we *feel* that we're getting the same findings, we will measure the resemblance of two heatmaps by their correlation. Actually, rather than using the correlation itself—usually referred to as *r*— we will look at the square, usually referred to as R^2. The benefit of using R^2 is that it has a simple interpretation: the percentage of variance explained. So, for example, if $R^2=.5$, then half of the variance is explained and the other half is due to random fluctuations.

By "explained" we mean that the data in one heatmap follow the data in the other, so that we can look at one heatmap and predict what the other will show. If R^2 was 1.0, we would have everything explained 100%, meaning that it wouldn't matter which of the two heatmaps we looked at—we could be 100% certain in predicting what the other one would look like. On the other hand, if R^2 was 0.0, there would be no correlation whatsoever between the two heatmaps and it would be completely

random whether the two had the same color at any given spot. (Yes, sometimes they would show the same color, but only as frequently as could be explained by random coincidence.)

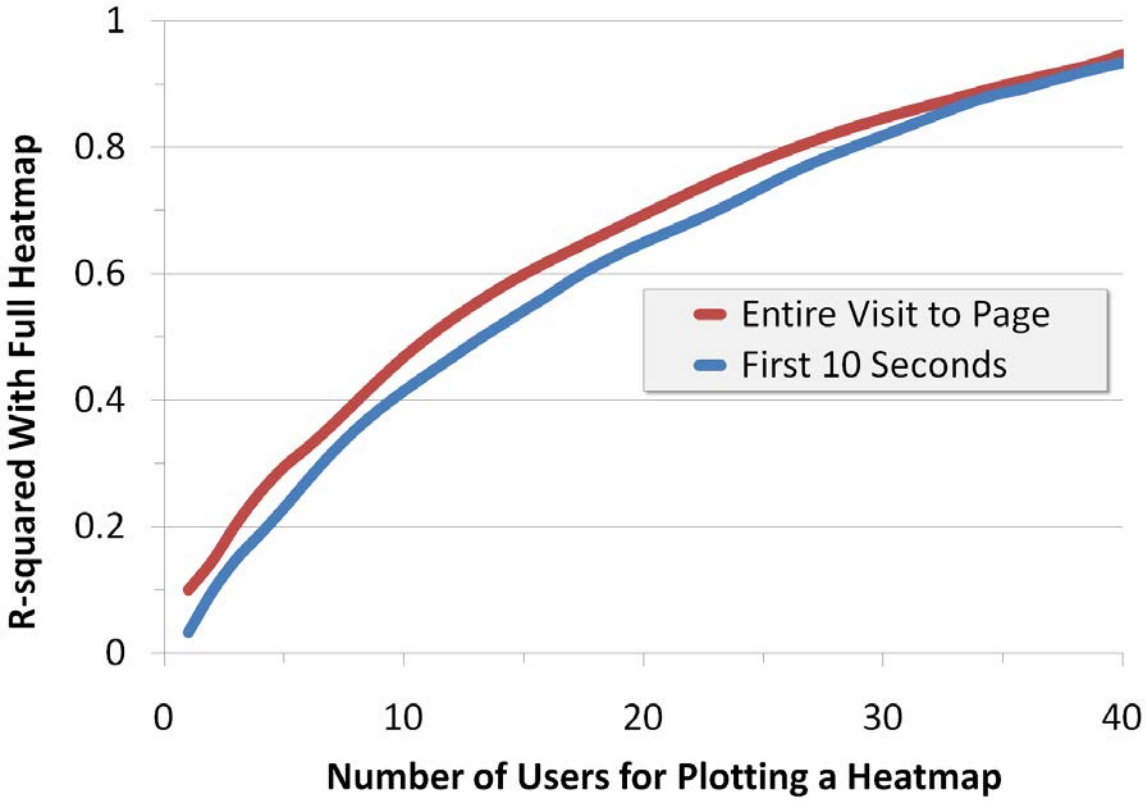
R^2 EXPLAINED

Let's consider an example of what R^2 means. We'll take two coins, say a dime and a quarter, and toss each of them a hundred times to record how often they come up heads. There ought not to be any correlation between how often a dime comes up heads and how often a quarter comes up heads. So if we repeat this experiment enough times, eventually we should record R^2 to be around zero. Knowing that the dime showed heads a lot in one round of the experiment should not give us any insight into the performance of the quarter in that round.

Our expectation of $R^2=0$ for games of heads-or-tails assumes that the coins are being tossed "fairly" by some reasonable meaning of that word. If the coins were being tossed by a professional magician who was a master at sleight of hand, it's quite possible that the outcome could be manipulated. Let's say that the magician had to decide before each round whether he or she was trying to maximize heads or maximize tails for both coins. Now the larger R^2 is, the better the magician was at cheating. Knowing that the dime came out heads a lot in a given round would make us believe that this was a round where the magician was trying to maximize heads, and thus we could safely predict a large number of heads for the quarter in that same round. And our prediction would be better (i.e., higher R^2) the more the magician could influence the outcome of each toss.

In eyetracking, the "magician" is the extent to which users tend to look at the same thing. If there's "strong magic," then we can safely predict all users' behavior based on the observation of a few users.

What we would like to do is to be able to run an eyetracking study with a fairly small number of users and use the heatmap from that study to predict what the heatmap would have looked like if we had tested many more users. In other words, it would be nice if R^2 were big when comparing heatmaps from a few users with heatmaps from many users.



How closely a heatmap plotted from a certain number of users will match the heatmap plotted from 60 users. Averaged across nine studies. The upper curve (red) shows the data for users' entire visit to a page and the lower curve (blue) shows the data for the first 10 seconds a user is on the page. "Closeness of match" is measured as R^2, which indicates the percentage of variance explained.

Unfortunately, as the chart shows, this is just not the case. R^2 is very small for the kinds of affordable studies with five users that we otherwise like to run in the usability field. The curves confirm the three conclusions we already got from qualitatively examining the three examples of heatmaps earlier in this section:

- The variability is higher for the users' initial exposure to the page than for their entire visit to the page. Higher randomness means lower R^2 for a given number of users—or, alternatively, that we need to test a few more users to achieve the same R^2 for the initial exposure as we got for the entire stay. The difference is small and virtually disappears once we test 35 or more users.
- With 10 or fewer users, the R^2 is too small for us to confidently draw conclusions purely on the basis of the heatmap. We might easily miss a major finding or—even worse—conclude something that isn't true. With R^2 at only .47, there's 53% randomness in the heatmap.
- Around 20 users, R^2 starts being big enough (about .70) that the main findings will be settled, but there's still a large amount of randomness left: around 30%.

- Around 30 users, the heatmap is pretty much nailed, with R^2 around .85, corresponding to only 15% randomness.

The conclusion from the curve is that we need to have eyetracking data from 30 users if we want to be able to draw sound conclusions from the resulting heatmap. In fact, if we want to draw conclusions from a heatmap of users' initial exposure to the page, then the heatmap should be plotted from eyetracking data involving 32 users.

Note that we need good-quality eyetracking results from 30 users. This means that we need to run about 39 users through the eyetracker to account for lost data with current eyetracking technology. (About 23% of users tested don't generate eyetracking data of sufficiently high quality, either because their eyes calibrate poorly or because the equipment acts up. Thus, if we test 39 users, we'll lose the data from 9 of them, leaving 30 good recordings.)

A BETTER DELIVERABLE THAN HEATMAPS

The solution to the eyetracking expense is to redefine the problem. Stop thinking of heatmaps as the deliverable from eyetracking just because they are the most colorful outcome and the one used for many eyetracking illustrations.

Yes, if your entire analysis is based on the heatmap and you want to draw conclusions based on the colors assigned to different parts of the page, then you do need to have data from 30 users behind the heatmap, meaning that you have to run about 39 users through your eyetracker to account for lost or poor data. So don't do that.

Instead draw your conclusions the old-fashioned way: watch what each individual user does as he or she performs your tasks and listen to what the users say as they think out loud. With this approach, the one difference between an eyetracking study and a traditional usability study becomes your ability to follow the user's gaze around the screen during the study and the ability to review the recordings in slow-motion to watch exactly where that user was looking. The slow-mo replays are often necessary because people scan Web pages so fast that you often can't follow exactly what words they read as they move through a set of pages.

You can still use the heatmaps, but only as illustrations, not as primary data. The primary data is the users' behavior and your analysis of what they are thinking, how they are solving the task, and how they are reacting to the various page elements. Once you have concluded that something is good or bad, then you can include the heatmap in your report or presentation to illustrate this point, as long as you don't make a big deal about whether a certain point is yellow or orange, because small color differences aren't going to be statistically significant.

If you're using the live test behavior supplemented by gaze replays as your primary data, then you only need to test with six users. Five would actually be enough to discover the vast majority of usability problems in the design, but you need that extra user to account for possible data loss due to poor eyetracking. Of course, with a small number of users, you might get unlucky and get low-quality recordings for two, or even three, users, but that's OK as long as you have taken good notes during the live testing.

The way to happiness (or at least a high ROI) is to conserve your budget and invest most of it in discount usability methods. Test a small number of users in each study and rely on qualitative analysis and your own insight instead of chasing overly expensive quantitative data. The money you save can be spent on running many

more studies. The two most fruitful things to test are your competitors' sites and more versions of your own site. Use iterative design to try out a bigger range of design possibilities, polishing the usability as you go, instead of blowing your entire budget on one big study.

Eyetracking Studies in a Lab

While an eyetracking monitor is heavier than most flat-screened monitors, it is arguably quite portable. However, once we got our labs set up, we certainly did not want to dismantle them and move them. Aside from this, there are various other reasons why eyetracking usability research, in particular, is often done in a lab setting. Incidentally, we use the term "lab" liberally, as a lab may not comprise the ubiquitous soundproof adjacent rooms complete with viewing monitors, great sound system, and a one-way mirror. Our lab, for example, was just a small office high in a skyscraper near Penn Station in New York City.

EYETRACKING LAB SET-UP NOTES

In addition to all the normal considerations when setting up any usability study, there are a few eyetracking-related things to consider when setting up your "lab" or users' workspace.

Low Light

Avoid spaces with windows or a strong light just over the eyetracking monitor. These conditions can hinder the calibration and tracking. Better to be in a windowless place (bordering on dismal that *Architectural Digest* wouldn't dream of considering) for eyetracking. Even though our lab was in a building with a fabulous view of the Statue of Liberty, we didn't pay the extra rent for one of these offices; for our needs, a windowless office was better, because it reduced glare and kept the users looking at the screen. We know, we didn't love this either when we were in the lab for months at a time. Seasonal Disorder sufferers beware.

Two issues with the eyetracking systems of the past that the Tobii system accommodates are *drift* and *head-motion compensation error.* Both could result in poor eye gaze capture. *Drift* is a weakening of a participant's eye calibration. It can happen due to environmental changes in the test lab, such as light conditions or humidity. For these and other reasons, the characteristics of the eye changes; for example, the pupil size may change or the eye may become dry. One way to avoid eye dryness is to limit the length of the test session to 90 minutes or less, allow the user breaks, and ask users to read tasks and answer questionnaires so their eyes get a break from looking at the monitor.

According to the Tobii reference documentation: "Drift effects for most users are reduced to less than 0.5 degrees, but may vary from individual to individual."

9. **Choose an office with good lighting, but not very bright rooms. Too much light can impact the eyetracker.**

User's Chair Should Be Stationary

The term *head-motion compensation error* is the additional error that can occur if the subject moves or turns the head. The field of view of the camera is about 20 × 15 × 20 cm (width × height × depth) at 60 cm from the screen. (This equals 8×6×8 inches at a distance of 24 inches from the screen.) Since it is enough that one of the eyes is within the field of view, this gives an effective tolerance of head-motion of about 30 × 15 × 20 cm. This is enough to compensate for any head positions that are comfortable when sitting in a normal posture in front of a computer screen. Tobii uses various patented techniques to accommodate this.

Give the user a chair that encourages him to sit still and not recline, bounce, or otherwise shift in any way. Ideally, get a chair that is stationary, does not swivel or roll. But, if you can find one, get a chair that does adjust vertically to accommodate varying user heights. The eyetracker monitor's angle feature also takes care of this.

10. **Seat the user in a stationary chair with no wheels or leaning or swivel capabilities.**

A Clear Workspace for the User

Unless you are asking the user to write something, consider removing paper and pens from his work area. These items often encourage users to doodle notes. We allow this in typical usability tests, but in eyetracking tests these items are discouraged because they may promote movement that could throw off the calibration.

11. **Remove pads of paper and pens from the user's desk area, unless you need them for your tasks.**

Dual-screen Set-up

We generally recommend that the facilitator sit with the user in the testing. (Note that some set-up needs to be done to accommodate more than two or so additional observers.) We think this makes the user more comfortable than talking with them through a booming "voice-of-God" speaker in the ceiling. And when facilitating, we typically look at the same monitor the user is looking at, sitting relatively close to the person but not uncomfortably so. In eyetracking sessions though, if you sit right next to the user you risk him moving to the side some or just shifting more during the tests. This could throw off the eyetracking. Also, watching the eye moving around the page really warrants its own PC monitor for the facilitator. These considerations and the fact that the facilitator needs to do several things between launching and saving eyetracking data make a dual-screen set-up a logical choice.

Doing website eyetracking testing is probably most efficient in a lab setting. Imagine trucking to a user's site with the eyetracking monitor in its case (one of those big, insulated heavy metal ones) only to find that there is too much light at the user's site, his chair swivels, or his eye cannot be calibrated.

Our colleague, Dr. Peter Brawn, usability consultant and eye tracking specialist at eyetracker, an Australian-based organization, has been working with eyetracking for 10 years in academic and commercial settings.

Peter's work demonstrates that the advancements in eyetracking technology are not limited to only stationary hardware. According to Peter, "Up until recently, tetherless eyetrackers required participants to wear obtrusive optics and recording units. However, eyetrackers are now available where the optics have been reduced so that they can fit on a pair of lightweight glasses, with recording units being modified camcorders or even transmit data to remote storage." He adds, "A good example is ASL's 'Mobile Eye' eyetracker."

In addition to the typical usability field study challenges, such as getting permission to do the research and later dealing with more data than you can imagine, Peter tells us about another challenge specific to eyetracking field studies.

"Clearly, when testing in situ, such as in a supermarket, the experimenter is no longer in control of the environment. You can't ask the store manager to dim the lights or move aisles around." Environmental differences in the field, for example, if

your study involved participants going from outdoors to indoors could pose challenges. He adds, “Therefore your eyetracker must be sufficiently robust so that it can handle different conditions without introducing inaccuracies.”

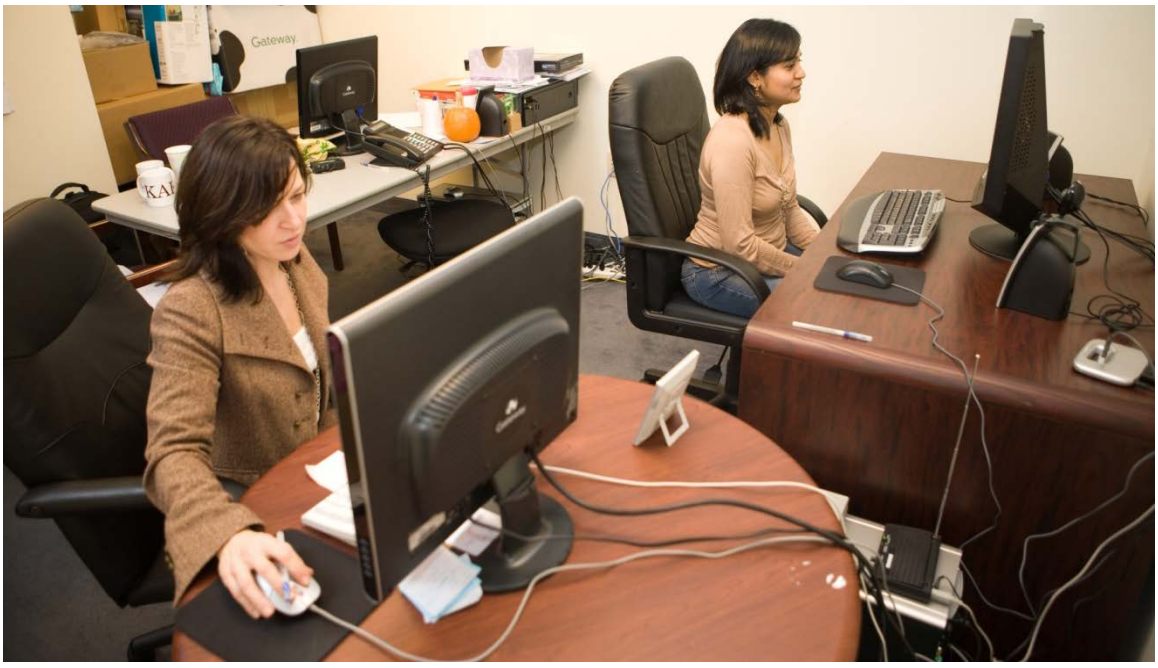
Interestingly, the technology itself does not introduce the biggest concerns in eyetracking usability testing in the field, Peter tells us. “From the perspective of capturing accurate data, the challenges with this type of testing revolve more around ensuring the optics are stable. People’s faces come in all sorts of shapes and sizes and spectacles have a tendency to slip down people’s noses. So you have to be able to adjust for this, as this will alter the pupil/bright spot readings and so introduce inaccuracies.”

And glasses slipping down one’s face cannot be comfortable. Peter continues, “As with any eyetracking study, another challenge is ensuring that participants are comfortable so that they will behave as naturally as possible. Like when you first put on a new pair of glasses, eyetracking optics can feel a bit strange, but after a few test trials people tend to forget they’re wearing them.”

A typical ET lab setup.

12. **In the lab, sit next to the user, but just slightly behind him so as not to encourage conversation. Watch his eye on a separate monitor using a dual-monitor set-up. Ensure that the facilitator’s monitor is not in the user’s line of vision. (Yes, if the user looks away from his monitor, the blue dot will disappear on the observer monitor anyway. But peripherally, the user shouldn’t be constantly aware of movement on the other monitor.)**

The facilitator sits near the user but slightly behind.



The facilitator watches on a secondary monitor.

Think Aloud Method

The think aloud protocol is the method we often use in our research studies. The premise is that users keep up a running commentary to say what they are thinking as they attempt their tasks.

BENEFITS TO THINK ALOUD IN A NON-EYETRACKING STUDY

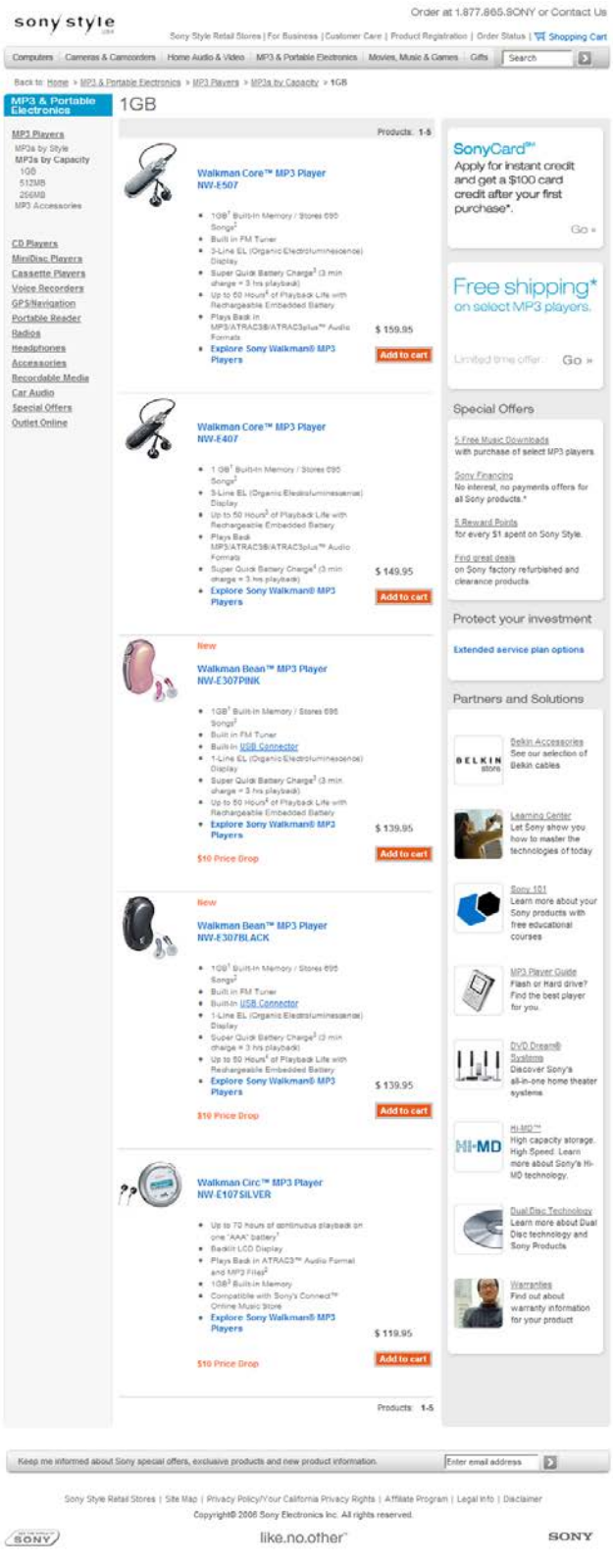
- It enables the facilitator and other observers to know what task the user is trying to accomplish. One might argue that merely giving a user a task to do in a usability study is what enables people to know what the users are doing. But this really isn't the case because people get sidetracked or even forget their task sometimes. Once when we tested an e-mail product, a user was asked to find a message that he had deleted. We thought the "deleted items" folder was difficult to find and see, and when doing the task, the user kept going to the other folders. At one point he said something to the effect of he knew where the deleted items folder was, but he was interested in looking at the other options. This, of course, is hard to evaluate since the user had not actually found the folder we wanted him to. But had he not said that, we would have continued to completely believe that he was looking for the *deleted items* folder the entire time.

- It enables the facilitator and observers to hear what the user is thinking as he does the task. Even if a user is not superb at articulating what he thinks, having him say something like, "What?" or "Oh, good" can help enormously in knowing what his reaction to the processes are.

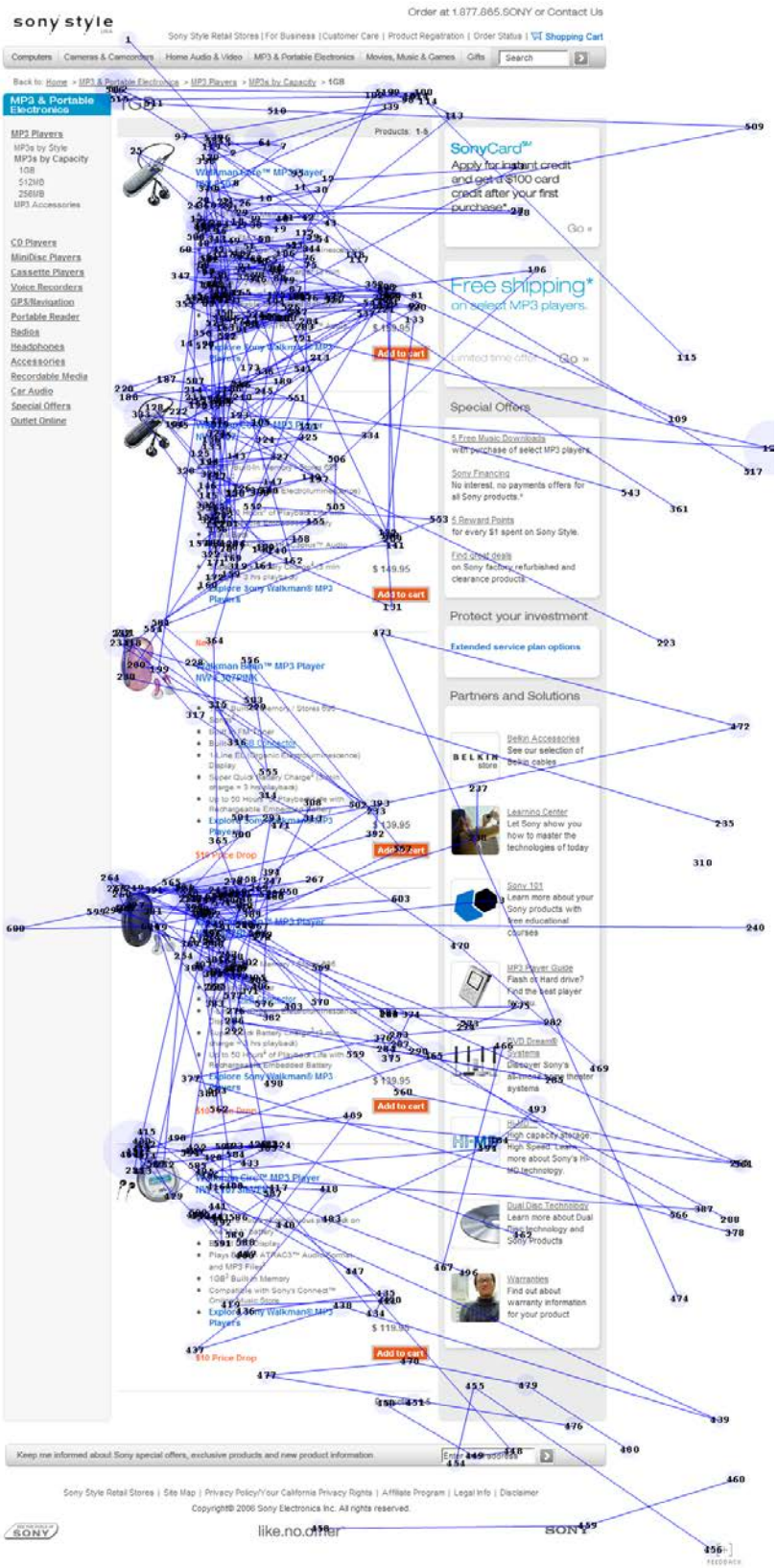
- The user's quotes and discussions can be entertaining and have impact on the development team. This may seem trivial, but the colorful imagery some people use can go a long way. It makes it easier to relate to users when you hear a bit of their feelings and personalities come out as they work with the design.

Ironically, in some cases the think aloud method can almost replace eyetracking technology altogether. If a user talks about what he is looking at it can be enough. And in some cases users do move their mouse to places where they are looking and are about to read. In one example when a person was trying to buy an MP3 player on the Sony website, he got a page that displayed only five players. He was surprised at this and said, "I would think Sony had more than that. But I don't see any *page one* or *next page,* which is usually here." The user then moved his mouse over the bottom right of the page, where he was also actually looking. He told us that he was looking there in the bottom right for a link that would display more MP3s. But there was none. He then scrolled up the page and moved the mouse over the choices, saying, "So, my only five choices seem to be these right here."

Following the mouse is the most low-tech eyetracking usability technology around. But we cannot rely on all users behaving this way during a study.



Sony's category page for MP3 players with a certain capacity, on www.sonystyle.com.



Caption: The user moved the mouse pointer around the page as he was looking for ways to find more players and as he considered each of the players in the list. The many fixations on the right edge of the page are due to the user looking at the scrollbar (which is not included in this gazeplot). In this case, a plot tracking mouse movements would have looked about the same as our gazeplot.

DRAWBACKS TO USING THE THINK ALOUD METHOD IN A NON-EYETRACKING STUDY

There are some obvious drawbacks to using think aloud even without eyetracking, including:

- Users may consider something more carefully as they try to articulate their thoughts about it. Say a person is somewhat attracted to an image on a page, but not really very drawn to it. They may glance at it, but when they start to make a comment about it decided they will look a little harder. And they may notice more about the contents of it as they are talking about it.

- The talking may hinder their thought process and detract from their focus on the tasks and the design.

- The additional task of talking about what a user is doing can affect quantitative objective methods, such as task time and success. Thus, when doing studies where we are concerned about these quantitative measures, we do not ask users to think out loud. Instead, they do their tasks in silence and are interviewed after the fact.

DRAWBACKS TO USING THE THINK ALOUD METHOD IN AN EYETRACKING STUDY

Now consider eyetracking—The point of doing the eyetracking is to get an accurate read on what the users tend to look at. Sounds simple, but in any usability study even the most expert facilitators can accidentally direct users in ways that they would not go if they were not being talked to or given tasks. In a typical study we know this can happen, but the effect these issues may have on the user is minimal compared to the effect they can have in eyetracking. The reason is because eyetracking findings are so specific. Say you are doing a typical non-eyetracking usability test and one of the things you want to know is whether people can create an e-mail message. Imagine that you are not sure if the button is in a location that is visible enough and you wonder if the words used are explicit enough. If you do not know where the user is looking, and they are trying to create an e-mail but they do not click the “Message” button, all you know is they did not click it. This is certainly valuable information. But you have to wonder why they did not click it. Two main reasons come to mind, one being they did not see the button at all. Or two, they did look right at the button, but the word *Message* wasn’t enough to make them think that this is the button to click to create a new e-mail. (Maybe they don’t think of mail as messages; something you might discover through think-aloud, which would also be a way to get inspiration for alternative labels for this command.)

Now imagine the same thing happens but you are using eyetracking technology. You would see whether the user looked at the button or not. If he didn’t even gaze over to it, the layout and look and placement are likely an issue. If he did look at it and read the word, the terminology is probably the culprit.

But if that study was done in such a way that the user was influenced by something, like a question from the facilitator, he might look at something he otherwise would not. Say the real problem with the button to create a new e-mail is the placement and layout. Maybe it’s in a spot too offset from the rest of the controls and it doesn’t look much like a button. Imagine the user has not even looked at it, but the facilitator prompts the user in some way, or the user’s own thinking aloud prompts

him to look more around the page. He then looks at the button. The eyetracker catches that the user looked at it, whether he clicks it or not, and the facilitator and observers may conclude that the layout is not the problem but the words are.

Now imagine the importance of these types of conclusions and acting on them. Of course in an iterative design process, you'd probably change the placement or name of the button, whatever the case warranted, and test it again. But what if you didn't? Either way, we should try to do the best tests and analysis as early as possible.

The main drawbacks to using the think aloud method in eyetracking research include:

- As a user talks about an item, he tends to look at it longer and more. Thus, a screen element may be burning red in the heatmap, simply because the user talked a lot about that element—maybe because it's difficult to understand—and not because the user necessarily engaged with the element during real use.

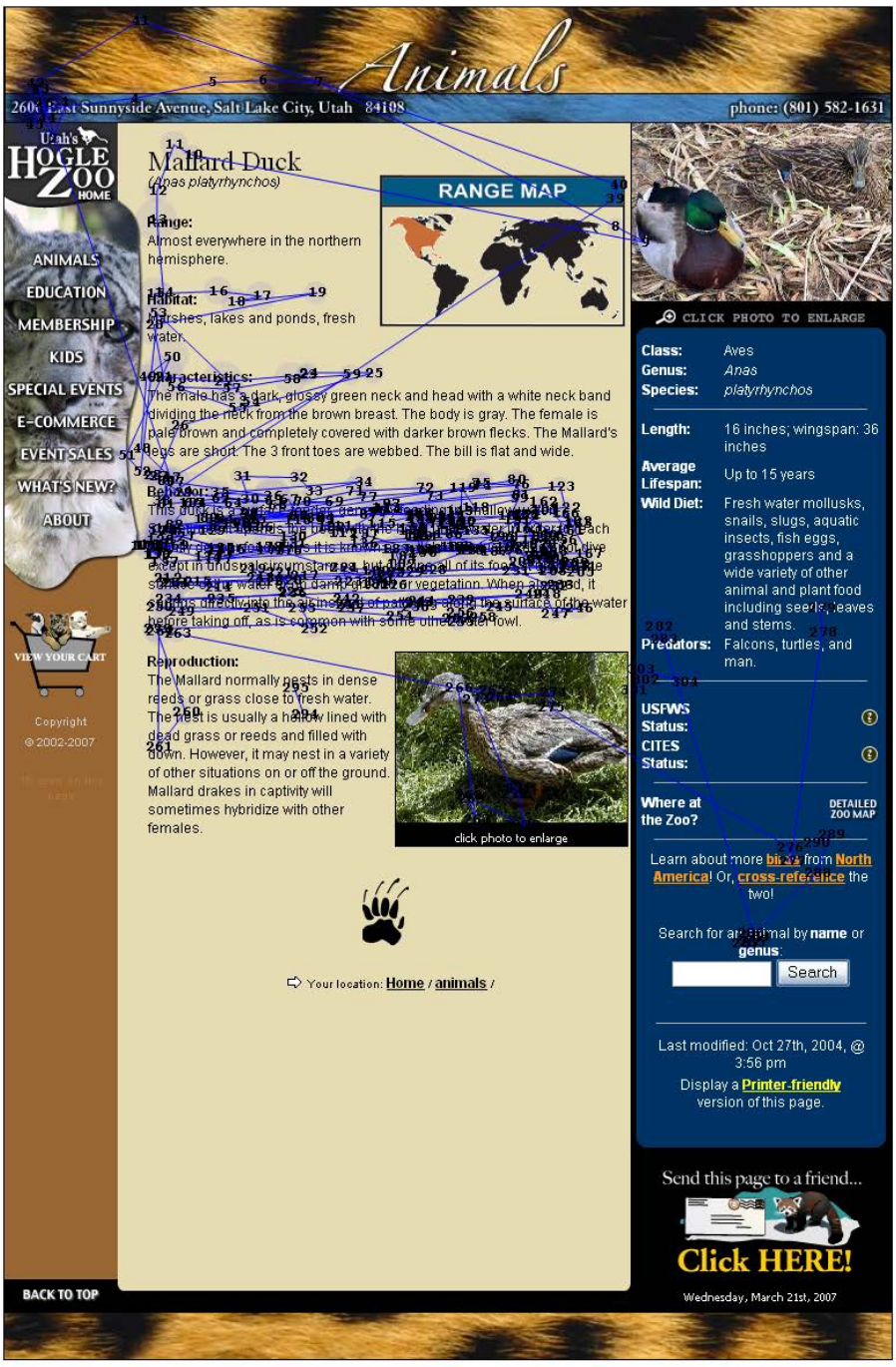
- As the user is speaking about a topic, he will likely look around more on the page he is currently on, even if he is not actually interested in or speaking about the items he is looking at. He simply looks around as he talks, practicing *perpetual viewing*.[3] This can make a page or item have more fixations than it actually would have had in practice. So some pages get inflated heat.

    - Or, he may notice and fixate on an item that he wouldn't have otherwise bothered with if he weren't doing the out-loud monologue.

In one example from our studies where a user was trying to learn about whether mallard ducks dive for food (they don't typically), she looked very differently at the same page when she was talking about it versus when she was just using it to do the task at hand. When trying to do the task, she was looking for something specific and was driven. She scanned the headings and wasted little time and fixations on words that did not have to do with her task. And when she found the heading most related to what she was looking for, *Behavior,* she read the text under that heading thoroughly. She gave one fixation to the image on the right of the duck. (Notice that she does not look at the second duck in the image on the right, maybe because it is camouflaged in the nest.) She looks a bit more at the picture of the duck toward the bottom of the page. She doesn't bother much with the map, and she looks minimally at the last paragraph and table on the right. She doesn't look at the left-side menu at all.

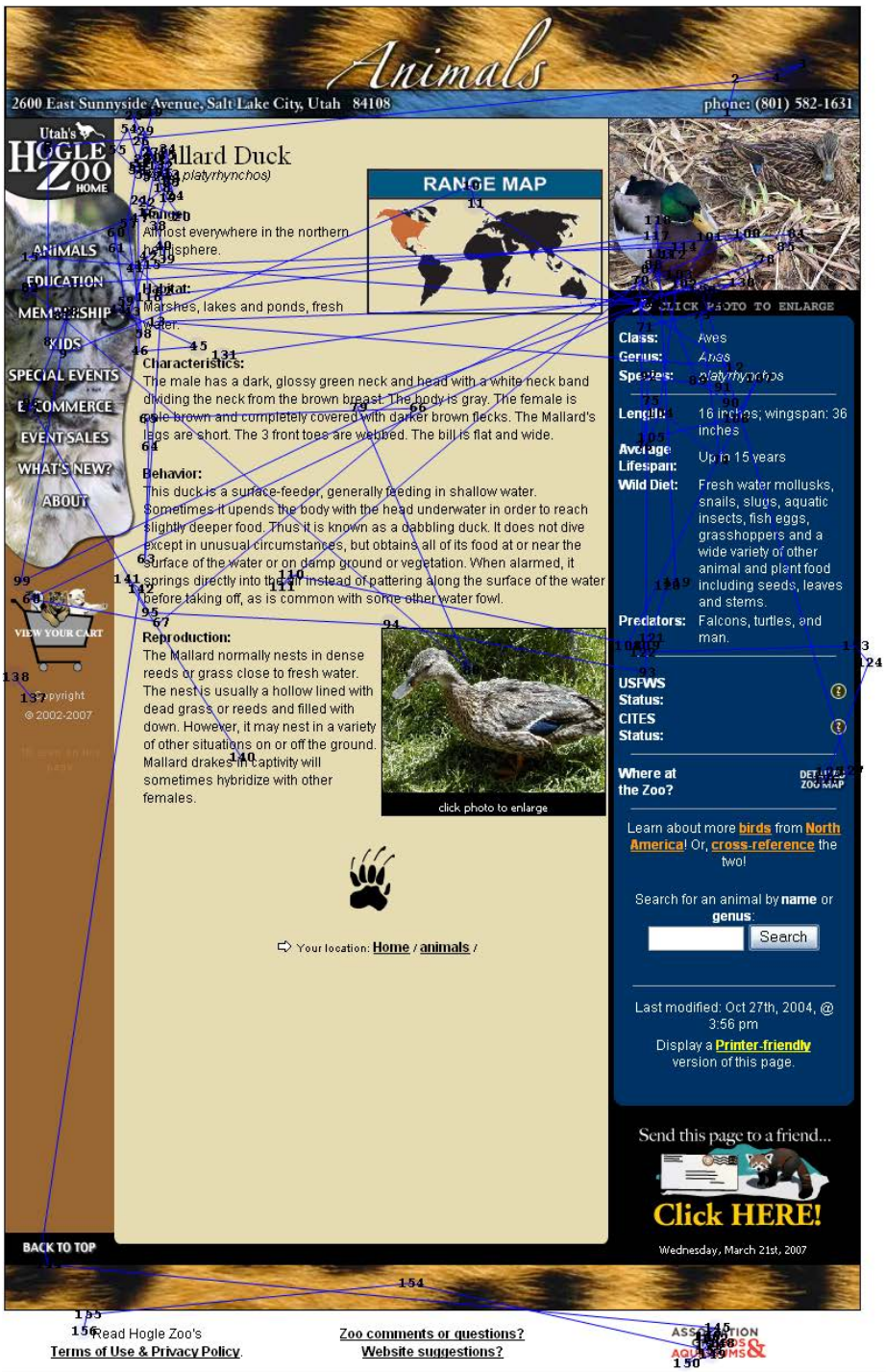
She did find this page via the Google search engine, and read the description there, so once on the page, she didn't really need the menu. And she found her answer in the main content area.

[3] One of the user behaviors discussed in further detail in our *Eyetracking Web Usability* book.



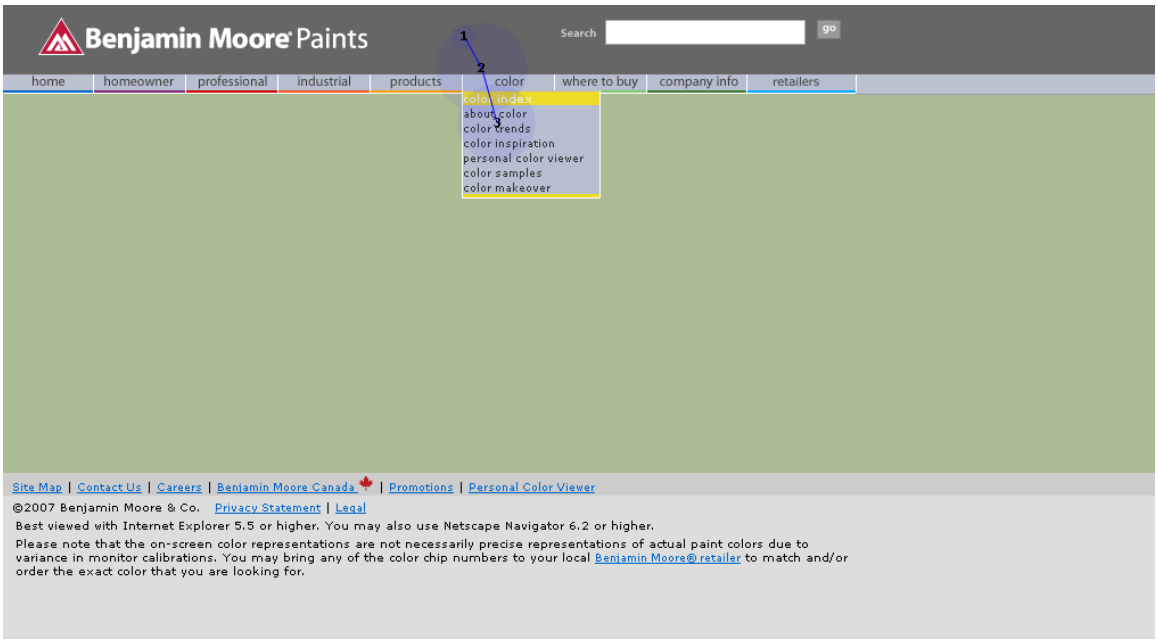
The user reads the text about mallard ducks when doing a task using www.hoglezoo.org.

But when this same user was asked to discuss the page, she looked at very different things, namely, a lot more at the image on the right, and at the left-side menu. She looked a lot less at the text and headings.

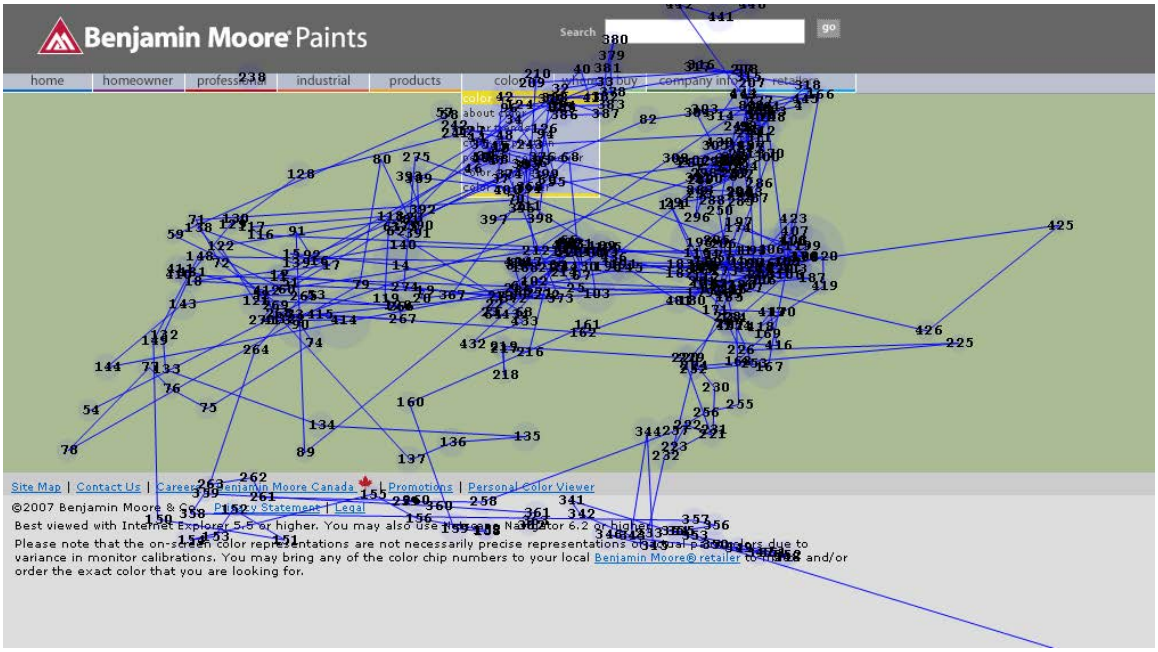


The same user reads less of the text in the middle and more menus and links during an interview, on a page on www.hoglezoo.org.

In another example, the user looks at a page on the Benjamin Moore website one way when doing a task, and a completely different way during a post-task interview. Note that the seemingly blank area was filled with dynamic graphics that were not captured in the screenshots that underlie the gazeplots. For more on such *misrepresentations* in eyetracking visualizations, see page 124.

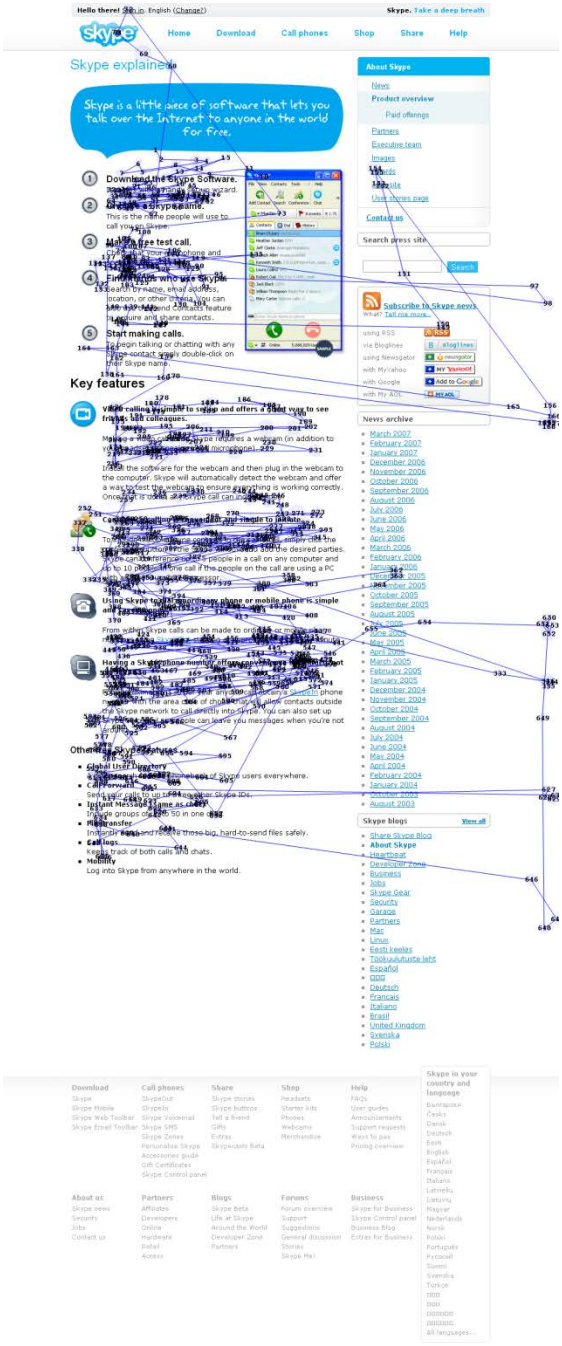


The user looks at the menus when doing a task on www.benjaminmoore.com.



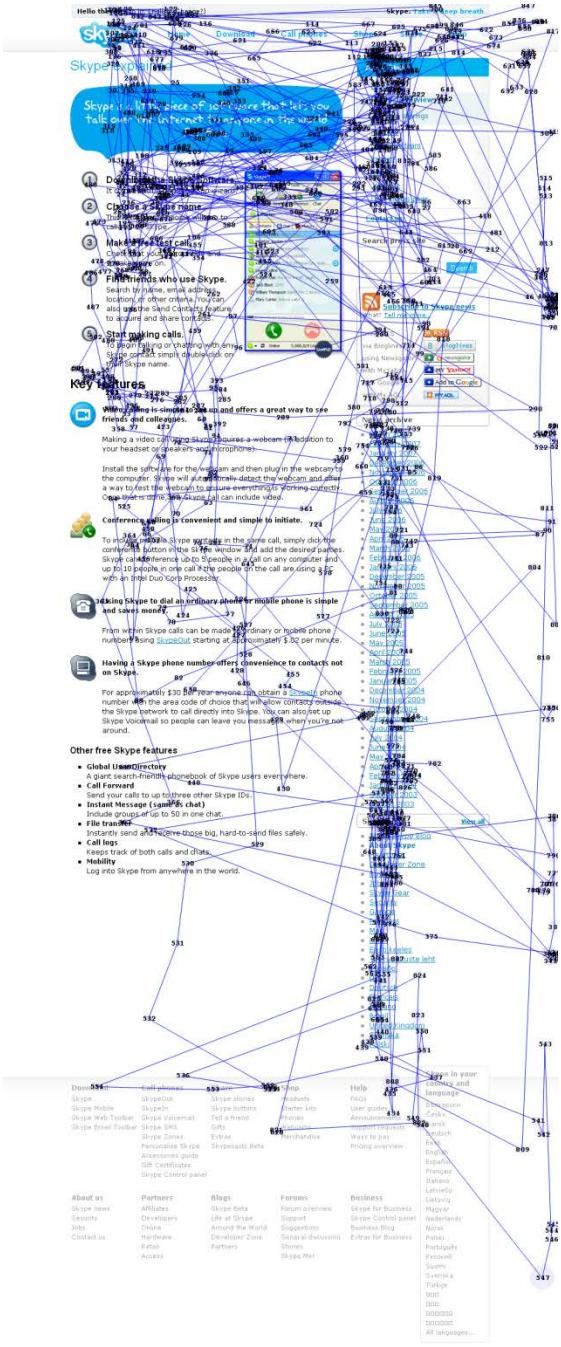
The same user looks at much more, including dynamic elements, during an interview after doing a task on www.benjaminmoore.com.

Another user reads about Skype when doing a task.



A user reads the text about the product on www.skype.com.

But during the interview she looks more at the links on the right.



The same user reads more of the links and menus during a post-task interview.

These examples demonstrate how the same user can look at the same page differently depending on the task, motivation, whether thinking aloud or silent, and whether it is his or her first or a subsequent visit to a page. With this in mind, of course, what users look at during a post-task interview is likely very different from what they would look at doing a think aloud task. The images above compare the same user 1) doing a task on a page, and 2) being interviewed about the experience. We believe these give a feeling, however, for some of the differences we would see if the user was thinking aloud while doing the task, the main point being that

researchers could get a very different read on how a person looks at a Webpage if he is thinking aloud versus not thinking aloud.

You can deal with these problems with the heatmaps if you recognize that the heatmap should be used to help you interpret the study as opposed to being the primary goal of the study. The purpose of data is insight, not the data itself, but if you want a pure heatmap, you should refrain from asking users to think out loud. For the vast majority of studies, the added insight you gain from knowing the users' thought processes is more important than whether you would have to make an allowance for spurious gaze durations during heatmap analysis.

13. **To generate pure heatmaps and gazeplots, do not ask users to think aloud.**

SURVEYING IS COMMON DURING THINK ALOUD

14. **Keep an eye out for *surveying* behavior in think aloud, where people appraise the page before they really try to do the actual task work. This may help them find features that they wouldn't have had they sprung directly into doing the task.**

One behavior we see people exhibit when they are talking about a Webpage during think aloud is **surveying and appraising the whole page.** Users will do a scan of much of the page, paying more attention to all of the main elements than they would if they were simply doing a task and not thinking aloud. Notice in the previous duck example that the user looked all over the page very thoroughly when she was discussing the page, but she spent the most time and fixations in the areas of interest when she was task-focused. Why is this? We have a few theories.

One possible reason the user surveys the page when he is thinking aloud is because he feels he needs to explain to the facilitator and observers, so he feels it is his responsibility to take note of everything available on the page and explain it to the onlookers. This behavior occurs more often earlier in the testing when the user is just getting acclimated. This is when we hear him say things like, "Okay, this is a nice page. I like it. I see the search over here and the pictures. There are the links. I like the colors." But, once this same users gets warmed up and becomes very focused on the task he is trying to do, he talks less and soon cares little about being a good tour guide.

The surveying phenomenon is one reason many studies sponsored by advertising agencies have bogus conclusions. By now there have been many reports published by various companies with a financial interest in increasing companies' advertising spend, and lo and behold, these studies always find that users pay a lot more attention to Internet advertising than what's found in studies not sponsored by the advertising industry.

We don't accuse the advertising companies of cooking their data. They don't need to. All they need is to use a study methodology that causes people to pay more attention to the peripheral parts of Web pages than they normally would. In one report we read, users were asked to comment on the design of a Web page. They were not specifically directed to comment on the ads, so when people remembered some of the banners after the session, the report concluded that it had proved the effectiveness of online advertising.

In fact, what almost certainly happened is that people looked around the page much more than they would have normally. If you visit a newspaper site to find out the latest baseball scores, you're going to scan the main menu for the sports section and then you're going to scan the results for the outcome of games you care about. Most likely you won't have a single fixation within any of the banner ads on the pages you're passing through on your hunt for Red Sox results. Visiting those exact same pages with the goal of commenting on the design is a vastly different task. In the latter case, you probably *would* look at the ads, since they're part of the page layout.

A study doesn't even need to be as blatantly biased as the one we described here. Even a small bias in study methodology can cause a big difference in study outcome when we're talking about something that almost never happens under natural circumstances, such as looking at banners.

Another reason why users may survey a page when thinking aloud is that many people don't just talk about a subject before they know something about it. (We have all come across people in life who know they are experts on everything and feel the need to tell us so often. One friend aptly calls this "answer syndrome." Basically, if a question is asked, this person will answer it and do it confidently, even if he knows virtually nothing about the topic. It happens. But we digress.)

In usability studies most people don't want to just begin waxing on about something they know nothing about. So when they hit a page, they look around a bit to make themselves better-versed on the subject. You can certainly see why a person would want to do this. But what they don't know is that we want them to talk about their actions, without studying the page or even offering a critique of the page. We want them for example, to say something like, "I see this link for *media* and since I am looking for a press release, I am going to click it." Click. User looks around for the words *press release* and doesn't see them right away. He says, "Where are the press releases? All I see here are contact names for the PR people. That's odd." Those comments and actions would be the sign of a perfect test participant, really, because: 1) we know what he was looking for, 2) we have an idea why he clicked what he clicked, 3) we know his reaction to the page that appeared, and 4) he even made a comment that reinforced what we thought he was thinking. Fabulous for interpreting and reporting usability findings, but not all users are really like this, understandably. And they do survey pages before giving their account.

No matter what the reasons behind it, the fact that users survey pages is not a trivial point in usability research, and it can have great impact on the user's behavior, success, task time, and feeling about the design. So test facilitators should not encourage test participants to survey pages. Of course, a methodical user, such as a very, very new Web user will tend to survey pages before he or she clicks anything. They should not be stopped. For the rest of the users, they should be encouraged to use the Web as they normally would. The words and gestures test facilitators use to prepare the participant can have great impact on how he will proceed throughout the study. So getting the set-up right is imperative.

When we tell a user we want his opinion, there is sometimes something lost in the translation—a disconnect between what we as researchers and designers need to know and what users think they are supposed to do or tell us. Many facilitators will say, "Give us your opinion on things." Some test participants take this to mean "Please do an artistic critique of each page you see." You cannot blame a user for thinking this. Really, how many times does a person get asked to talk about the most mundane thoughts going through his head, like exactly how he thinks something might work.

A participant might believe that observers would be a lot more interested in his assessment of the aesthetics. For example, assume that opening a door is the user's task (a nod to our colleague Don Norman's infamous door example in his book *The Design of Everyday Things*). Walking up to that door, a user might think we want to hear, "Nice door. I like the glass. I like the chrome. Wide. Nice."

This is all fine commentary. But what would be more helpful to creating easy-to-use doors? Watching him use the door, and noticing whether he pauses to look at it, pushes when he should pull, or gropes for the handle. Notice whether people can get through it quickly and safely. And the accompanying commentary we want would be something more like, "I wonder if this door is going to push out or pull in. I think from the way the hinges are facing I should push. Here I go." The user pushes. Nothing happens. "Well, I guess you pull the bar then." The user pulls and the door opens. "Oh, I see. The hinges are over here. I don't know what those other chrome things on that side are then. That's embarrassing."

We can try blaming the test facilitator for soliciting these aesthetic responses. But it is often not her fault, for even the best facilitators who take all the right precautions, those who discuss what they need, give a demonstration, and even have the user practice, will still on occasion only get these artistic comments from users. It's close to impossible for some people to believe that observers actually care about their stream-of-consciousness. Not James Joyce fans, perhaps.

Some artistic comments can certainly inform designers. But true artistic critiques should come from a designer or an artist, not from a typical user. What we really need that user in the lab for is to watch him actually interact with the design and to hear how he feels about it in the context of use. Once a participant starts solving problems and offering solutions, the test is just about shot. There is a massive difference between a user saying:

"Oh, there's the *Add to cart* button. I couldn't find that. I was looking over here for it before." And:

"You should move the *Add to cart* button over to here. And make it bigger than the other buttons. Maybe use that same blue up here in the header. And I think a border would be really nice here . . ."

Some believe strongly in the Participatory Design method, where users and developers work together to derive a design. We agree that each and every usability and design method has its place and good reasons for using it under various circumstances. But calling on a user or potential user to design a typical Webpage is not the most efficient way to use either the participant's time or the developers' time. Basic economics logic tells us to use resources for what they are best for. Remember in Econ 101 when our professors showed us various examples of how many utils (bits of energy) farmers use to yield different crops? Assuming that all of the farmers need to produce something, the farmer who can use the least energy to harvest the most in relation to the other farmers, should farm that particular vegetable. For example, say Farmer A expends 100 utils to produce 50 squash. Farmer B expends 100 utils to produce 60 squash. Farmer A expends 100 utils (bits of energy) to produce 20 tomatoes. Farmer B expends 100 utils (bits of energy) to produce 100 tomatoes. Farmer B is clearly the better farmer (or at least has better land and equipment) but farmer A needs to farm too. While Farmer B would produce more squash for the same energy as Farmer A would use, Farmer A is a complete train-wreck trying to produce tomatoes. So let A do the squash and B do the tomatoes.

Similarly, some users will have some good design ideas, but those who are experienced in and/or trained in design will be far more efficient and should be the

ones doing the design. Let the user do the using, not the critiquing, as he can sometimes be a complete train-wreck when attempting design.

There is a great Simpsons' episode[4] where Homer's half-brother, Herb, an automaker (at the time, anyway, before Homer got involved), gave Homer, an arguably everyday man, complete design control over a new car design. The design team at first ignored the dopey suggestions he made, until Herb berated them and told them to do everything Homer said to do. So they obeyed the order and after huddling around Homer for a suitable period, unveiled the most hideous, useless, and expensive car at the car show, ultimately bankrupting Powell Motors and putting Herb on the street. Ah, we can learn so much about product development from the Simpsons.

Let's not put the user in this awkward position that beckons colossal failure anymore. Instead, there are some things we can do to properly guide the user in our behavioral research study. When doing think aloud testing, explain to the user:

- Tell us what you are doing and thinking as you go. We are interested in your feedback. What we really need is your thoughts as a person who is trying to use this. How is it? What's working and what's not? And keep us apprised of what you are looking for and trying to do.

- Don't ask a user how he would design something, and don't really encourage design if he offers on his own. If you write profusely and nod vigorously, it will encourage the person to continue to design. Better is nodding slightly and not writing. This usually does the trick. And it is not mean to demonstrably exhibit low interest in design suggestions as long as you don't actively reprimand a participant for making a design suggestion.

Give tasks, or let them choose their own. Of course, open exploration and completely open research in a field study, for example, is all very telling and we recommend it. But in eyetracking research, in particular, it is very helpful to know what the user is trying to do. And most importantly, we are encouraging a task, either given by you or created by the user, as opposed to saying, "Look at this one page or image. What do you think?" For if you do the latter, you will not get a good eyetracking read on how a user would actually traverse a page. You will get their looks as though they were told to spend time on and look at something, which is not realistic.

Of course users are smart and many even have some good design ideas. More to the point, for some people the only way they express their thoughts during a study is through design ideas. Hearing a user's design ideas can help a test observer understand the user's issue sometimes. For example, if a person says, "I think you should redesign this page as a three-column layout," this can indicate to observers, at the very least, that there is something wrong or bothersome with the current way the page is laid out. It may mean that the user thinks the page is too cluttered, and suggesting three columns is his way of cleaning things up. But really we do not know this based on the user's design suggestion. So if a user says something like, "I think you should redesign this page as a three-column layout," the facilitator should wait a few seconds to see if the user offers a reason for the suggestion. If he does not offer

[4] For copyright reasons, we can't show you Homer's car, but the unbelievably nerdy nature of Wikipedia comes to the rescue and tells us that this was episode 28 of the series, originally aired February 21, 1991 during Season Two, entitled *Oh Brother, Where Art Thou?* An image search for this title plus "Simpsons" and "car" should be enough to allow you to track down a picture of the user-designed car.

one, the test facilitator may say something along the lines of, “A three-column layout. . .?” and see if repeating part of the phrase the person used prompts him to explain why. If he does not say why (and instead explains what a three-column layout is) then the facilitator may (at the right time) ask something a bit more directive and intrusive, such as “Can you tell me why a three-column layout for this page?” or “Can you tell me what makes you want to change this page? Why?”

Beginning the Session and Calibrating the User's Eyes

There are many things to communicate to the participant at the start of a session. This is challenging because when you tell the user too many things, they sometimes forget what you said. Or a long laundry list can make a person nervous. But on the flipside, if you don't tell the user pertinent information, he will wonder why you are doing certain things and, again, become nervous or mistrust the facilitator. Or, for example, if you don't fully communicate that he must tell you when he has completed a task, then it is very difficult to judge the task timings and task success scores, and when to stop the tracking recording if the user is actually finished with the current problem but still looking around the page because he is waiting for a cue from you.

If you try to communicate these things during the session you may shock the user or make him feel he did something wrong and you are correcting him, or impact where or how long he looks at something. Whereas if you tell him before the start of the session, then when you "correct" him, his reaction will be more of, "Oh, right." He knew it was coming, it's not him, the facilitator expected it to happen, and it has happened to other testers before. Also, correcting the user during the study can impact his task time and success scores for the task you made the correction on.

Striking a balance between these issues is difficult to do. But after doing hundreds of eyetracking sessions we have a good base sitting somewhere between inundating the user and giving enough information to make the tests run well. The transcript follows, and some of the more important eyetracking-specific usability method points are annotated in bold italic so you as the reader know why these points are important.

WHAT TO SAY TO THE USER WHEN BEGINNING THE SESSIONS

Thank you for helping us with our study.

Before we begin, I'd like to review a few things with you.

We are trying to learn about what is easy and what is difficult about websites, to ultimately make them easier to use.

The only way we can do this is by watching people use them.

So, that is what we are going to do.

I will give you some scenarios and ask you to work through them.

About Scenarios/Process

Some scenarios are open-ended, and some are very specific and ask you to find an actual answer.

Work through the scenarios as best you can.

There is no right or wrong answer.

As you work, there is no need to tell us what you are doing. Just act as though you are at home or work.

When you have your answer, please tell me, "I have my answer."

It's very important that you tell me this, otherwise I will just keep writing and watching.

Also, if you think you will not find the answer and want to stop, or you just want to stop, just let me know.

So for example, let's say the scenario is to find the zodiac sign if your birthday is June 15 and to find the answer using a specific website. I will open the website for you. You will use it, and when you find your answer, you will say, "I have my answer."

At that point, I will do some things over here (gesture toward the facilitator's monitor), which will take control of your mouse and your screen. Otherwise, you will have total control of your mouse and screen.

We found that even with a dual-monitor set-up in the lab, the facilitator did need to take control of the mouse to stop the task from running and recording in the eyetracker. In our pilot studies we saw users a bit confused when their mouse stopped working and also very jarred when their Web screen just went away on them.

So, the website you were using will be gone.

And you will say "Gemini." Okay?

After Scenarios

After most scenarios, I will ask you some questions about how you felt about the website design.

And in a few cases, I will give you a questionnaire about the actual information you read, about what information you learned. I will tell you when I give you the scenario if I am going to do this.

Remember, even for those questions, this is not a test of you. It is helping us to learn about websites.

More Process

We will be watching what you are doing on this monitor.

Mostly, I will not talk with you as you are working.

If I talk with you I could affect what you do. (The observers won't be talking.)

So, it is most helpful if you work independently—that's how we learn the most.

If you have questions, you may ask them but I mostly cannot answer. If I know the answer, I will answer at the end of the session.

Forms

Let me give you your honorarium. Thank you for helping us.

> Give $100, count it out and say, "You can put this away."

This sheet acts as a receipt for the hundred dollars, and by signing it you acknowledge that you agree to participate, and we can use the data from the session in any reports, books, or presentations.

Any personal information you give us, here or during the session (like your address), will be kept completely confidential.

Please take a moment to read this, and if you agree, sign it.

> Give consent form.

Now we'd just like to get a sense of some of your interests. Please check any of the topics that you are interested in.

> Give the user the Interest questionnaire.

Do Workspace Set-up for ET Calibration

The eyetracking system needs to be calibrated to find the particular user's eye. This is done in the lab at the start of the session.

As mentioned previously, we will be videotaping the computer screen and where you look on the screen.

So, first we need to do a set-up item.

Do you mind if I clip this microphone on you?

> Clip microphone to collar, up close to mouth.

Now if you would, please move in your chair to a comfortable position and look at the monitor. I just need to get this workspace set up a bit, so I am going to stand behind you and make sure things are in place. Please just keep looking at the monitor.

15. **Inform the user of what you are doing, especially when you are standing behind her while you adjust her position and consult the facilitator monitor.**

The facilitator checks the user's position for optimal tracking.

> The user should appear in the middle of the monitor and the correct distance from it, based on the particular technology's specifications.

Note that this and the rest of the photos in this report showing user set-up for eyetracking studies were not taken during our actual studies, in order not to disturb our test participants. (The "user" in the photos is actually our research assistant, Sukeshini Grandhi, and the facilitator is Kara Pernice.)

There's a second reason to avoid showing photos of real users, besides the desire to avoid disruptions during the research. One of the main ethical imperatives in usability is to preserve the anonymity of test participants so that they can feel free to speak their mind without fear of retribution. Thus, we never report the names of our test users, and we also try to avoid showing their faces, as much as possible.

The user sits in the center of the monitor.

> Adjust the monitor angle/placement.

Excuse me, I am just going to adjust this a bit.

16. **Look for a solid "lock" before you calibrate the user. A lock occurs when you can clearly see both eyes in the middle of the window, and there is no excessive blinking or disappearing of either eye.**

> Check that two eyes show up on the facilitation monitor. A lock occurs when you can clearly see both eyes in the middle of the window, and there is no excessive blinking or disappearing of either eye. They should be in the middle of the screen. This point is basically the moment of truth—can I get this person's eyes?

The observer monitor displays two dots representing the eyes, indicating that the tracker is aware of the eyes.

This is also the point where you may need to adjust the monitor angle, monitor placement—left/right and closer/farther away—or the user's chair. Note that about 20 inches (50 cm) away is the typical recommended distance between the user and the monitor, but this also can depend on the user's height and the distance he is sitting from the monitor. In more extreme cases, we had to put a box under the monitor to get the angle right. Also, this is where you might have to ask a user if he would remind removing his hat or moving his glasses up or down his nose. This is a bit awkward, for sure, especially if you are alone with the user and you have to ask her to slide the box under the monitor while you grunt and lift it up. And that thing is heavier than it looks.

The user helps the facilitator to elevate the eyetracker by slipping the box under it while the facilitator lifts it. Note that the eyetracker is heavy, so watch your back if you are lifting one like this. And don't allow the user to lift it.

> If you cannot get a lock, try all of the adjustments. Say something to the user so they do not feel bad, like: Getting this technology exactly right takes a few tries. Sorry about that.

17. **Adjust the monitor, chair, whatever you need to do to get a lock. Don't be shy about it.**

If you don't get this right, you will probably have a wasted session—at least as far as eyetracking is concerned—so it's better to spend a few more minutes on set-up than to waste the entire 90 minutes you're paying the user to be in your study.

18. **Give the user verbal feedback as you try to get a lock, so they don't feel confused or like something is wrong with them. Say something like, "It takes a few tries to get the technology set up."**

> If it's impossible to get a lock, either release the user (and let him keep the money), or still do the usability study without the eyetracking element. If you do not know the cause of the poor calibration, you may try some gentle, tactful probing to figure out why the ET did not capture the eye, to learn for the future some additional questions to put in the screening questionnaire.

19. Do not calibrate the user until you are ready to start running the tasks.

The first time you ask the user to look at the eyetracker and you consult the facilitator's screen is the moment of truth. This is where you get a sense of whether this person's eyes can be tracked well or at all.

> Once you have the eye locked in, say:

I am going to open a new website. When I do, would you please use the site. Look for something of interest to you and read it. Go ahead and hold the mouse, use the keyboard, look at the monitor; just read anything you are interested while I adjust things around you. Just act as though I am not here. And when you feel finished with reading an item, please tell me you are finished.

This is not a task that we record. In fact, we couldn't if we wanted to because the eyetracker is not yet calibrated. Among ourselves, we call this a Practice Task. We never mention to the user that this is "practice" because hearing it is just practice might make them infer that they do not really need to fully do it. Also, later when we give them the "real" tasks they might get nervous.

20. Give users a task with a website before you calibrate, so they actually get into the position they sit in to *use* the Web.

Having the participant actually use a website serves a dual purpose. First, it gets the user to put himself in the right physical position that he would sit in to click, scroll, and read on the Web. In our pilot studies we found that when we did not do this and instead we just encouraged the user to get into position and said, "Sit as you would if you were using the Web. Use the mouse, move your chair, use the keyboard," the user really thought he was in position. Then we calibrated him. Then we gave him his first task. What was the first thing he did now that he was actually trying to read and click? He moved his chair and position in that chair, thus often completely throwing off the calibration we just carefully did. So, we can throw away task one. (And we don't like to do this as throwing away eyetracking data quickly gets expensive.)

The second reason for giving a Practice Task is to get the person used to telling the facilitator when he is finished with the task. This is something we do in our non-eyetracking quantitative studies as well. But it is especially important to do in eyetracking studies to ensure that once the user is done with that task that fixations are no longer recorded. This makes using the heatmaps and gazeplots easier to deal with. If there were a lot of extra time on the recorded eyetracking tasks, it would be more difficult to make accurate heatmaps. During the Practice Task, users who do say they are finished are thanked for doing this and told that this is exactly what we need. Users who do not

tell the facilitator they are finished are reminded to please try and do this.

21. **Give the participant a "Practice Task" before you try and calibrate his eyes. Ask him, for example, to read a news item, or to do some other real thing using the Web *before* you calibrate him. This gets him into the position he will actually want to sit in to do work far better than just asking him to "get comfortable" and "hold the mouse" does. Once he is settled in, calibrate him. Then give the first task.**

Do ET Calibration

Okay, please continue to look at the monitor.

I am going to close the news site you are looking at, and you'll lose control of your mouse. And in a moment you will see a blue dot on the screen. It will move around and get larger and smaller. Please **watch** the blue dot as it moves.

> Calibrate.

> Review the calibration status. All of the points should be as close and tight in the circles as possible.

After calibration the eyetracker shows the facilitator only which look-points the tracker was able to capture and calibrate.

We'll do this a few times. Here comes the blue dot.

Here we want to communicate exactly what is going to happen, what they should do, and that we will be doing this a few times. Warn the user the dot is coming or this can be jarring. Setting the user's expectations without making a big deal about everything makes things run easily, smoothly, and correctly.

The user watches the blue dot as it moves around the screen.

The user watches the blue dot as it continues to move around the screen.

The user still watches the blue dot as it continues to move around the screen.

The user watches the blue dot as it further moves around the screen.

The user watches the blue dot as it hits its final positions on the screen.

22. **Before you invoke the calibration, warn the user that the website will go away and a blue dot will appear.**

23. **Do not say things like, "We have to do that again" when calibrating. People think there is something wrong with them if you say things like "that didn't work" or "we'll have to do that again." If you are more carefree in your voice and words, the user will remain comfortable.**

Do this as many times as needed—at least two or three to check for consistency. Choose the best.

Do NOT say: We need to do that again.

Rather, DO say something light and non-accusatory, such as: And we'll do that again . . .

That's all for the blue dot for now. We'll probably do that a few times again later.

We want to prepare them for later in the study when they move or their eyes get a little mal-tracked and we need calibrate again.

Look at the Monitor

It will help if you look toward the monitor while you are working. I know it is odd, but even if you are talking to me, please look straight ahead.

Most users will start staring at the monitor now, as you say this. That's good.

And please try not to move around too much. You certainly don't have to sit like a stone. You can move your head, just not too much.

Also, please try to keep a clear path between you and the monitor. Most people at some point lean on their hand or something. Please try to avoid this as it can interfere with our equipment. But if you do it I'll just remind you to move it. Okay?

You have to tell the user to sit relatively still and warn them that you are going to tell them to move their hands if they are in the way. Again, it is so important to set expectations so you are not trying to fix a problem with these things during the actual study.

24. **Tell the user to look toward the monitor, to not move around much if he can help it, and to keep a clear path between himself and the screen.**

Final Reminders

You can take a break or stop working on a scenario if you want, just let me know.

Again, remember that this is in no way a test of you or your ability. Rather, you are helping us evaluate the designs.

Again, when you are finished with a scenario please tell me when you have your answer.

Do you have any questions before we begin?

25. **Remind the user to tell you when he is finished with the task at hand.**

Begin

> Give first scenario.

26. **If giving a printed task, wheel your chair or walk over to the user. Do not make him move at all from his well-calibrated position to reach anything.**

The facilitator scoots over to the user to hand her a task. The user should move as little as possible once the eyes are calibrated.

The facilitator scoots over to the user to hand her a task. The user should move as little as possible once the eyes are calibrated.

Impossible Eyetracker Calibration

Ideally, you will screen for and rule out the most impossible-to-calibrate situations. After all, if the eyetracker cannot find the user's eyes, what is the point of having the user come to the lab at all? In our case, we ended up still doing some of the tests where the users couldn't be calibrated, just to collect more test ideas and background info. But if you are looking for data about the eye, the system has to work. And there are some known factors that make it difficult if not impossible to calibrate some people's eyes.

Basically anything that gets between the eyetracker's illuminator and imaging device and the user's pupil can hinder calibration. The following are some of the offenders.

CANNOT BE AMENDED: THESE CULPRITS REALLY CANNOT BE CHANGED TO MAKE CALIBRATION POSSIBLE

Bifocals

Bifocals and trifocals usually cannot be calibrated. Progressive lenses are basically a form of bifocal or trifocal. But we found that some people do not think of their regression lenses as "bifocals." In our original screening questions we asked users whether they wore bifocals. When a few people arrived at the lab wearing glasses and could not be calibrated, we then found that the glasses had multiple prescriptions. The users just never heard them called "bifocals" so did not think it mattered. So make sure in screening that you ask the questions about glasses to solicit whether the person has multiple prescriptions in their lenses. If they do, the calibration usually does not work.

Thick-rimmed Glasses

Ever since the first American made his way over to Germany, we have been wearing cooler and cooler glasses. Andy Warhol, step aside. These daring colors and ultra-thick rims are fabulous for making a bold fashion statement but are terrible for eyetracking.

Permanently Dilated Pupil

On a few occasions we could only track one of the user's eyes. Try as we may to adjust the monitor, the chair, the desk, and recalibrate, the tracker would only find one eye. As it turns out, some users had one pupil that was permanently dilated, which cannot be tracked. (They share this affliction with the great David Bowie who was punched in the eye in a fight over a girl when he was a teenager. So if Bowie volunteers for a session, keep him in your database for a different study.)

Glaucoma and Cataracts

These were other afflictions that were not possible to calibrate, or that gave potentially unsound reads.

User is Extremely Tall or Small

Sometimes the user is so tall or so small that angling, lifting or otherwise moving the monitor or chair will just not make it work. This is rare, but it has happened.

User Has Very Long, Thick Eyelashes or Is Wearing a lot of Mascara

When you see a person walk into the lab with enviable eyelashes, you know you are in trouble. These thick babies can actually impair the eyetracking. And when some people wear false eyelashes, or believe it or not, mascara like Tammy Faye Bakker, you may not be able to capture their eyes.

CAN BE AMENDED: THESE CULPRITS USUALLY CAN BE CHANGED TO MAKE CALIBRATION POSSIBLE

An ideal participant in an eyetracking study:

- has no glasses
- no hat
- no physical objects that could potentially get between the eye and the eyetracker
- sits relatively straight in the chair
- does not fidget or move too much
- remains at about 20 to 23 inches from the monitor at all times (check your particular technology for the appropriate distance)
- remains positioned at the middle of the monitor at all times.

It sounds like prison, but it is really possible and not uncomfortable as long as test sessions are limited to less than two hours and you have a comfortable chair for the users to perch in.

The user is in a good position for the eyetracker.

Now let's look at many of the situations that arise in doing testing with eyetracking.

Brimmed Hats

Some users show up wearing a hat. The brim sometimes gets in the way and impairs calibration.

A baseball hat may impair the eyetracking.

Even a short brim may impair the eyetracking.

We usually attempt to calibrate the user with the hat on, as we found asking a person to take a hat off sometimes made them self-conscious. If the calibration did not work, however, we did have to ask them this. We typically said something like this: "I am sorry. It seems that your hat might be interacting with our system. I wonder how you would feel about removing it?"

27. **Try to calibrate the person if he is wearing a hat. If it does not work, gently ask him if he would mind removing the hat. If he's uncomfortable, do not force it.**

Looking Over Eyeglasses

With thicker-rimmed glasses, or even some that don't have terribly thick rims, if the user moves the glasses down and looks over the lenses, sometimes the glasses interfere with the eyetracker.

Glasses sliding down the nose may impair the eyetracking.

Hair Bangs/Fringe

In some cases when people have a lot of hair on the forehead, it can get in the way of the eyetracker.

Drinking

We always offer the user a drink at the start of the session, especially if we are asking the user to talk a lot. But in the eyetracking sessions, when they hold the cup up while taking a sip, it often gets in the way of the eyetracking. This is usually only for a few moments. Thankfully, the eyetracker just drops the eye and doesn't give some invalid read.

Giving the user a drink, even just a glass of water, at the beginning of a session can help them to feel a bit at ease. Also, if they are working in the lab, especially if they are thinking aloud for more than an hour, they probably will need some beverage to wet their whistle. But don't offer people food in the lab. It is often very distracting and can take time away from the session.

Drinking from a cup may impair the eyetracking.

In the above image, the cup is in the way, plus the user has long bangs that might get in the way of tracking.

Holding Cup Up

Of course, once the user takes a sip, she doesn't always just put the cup down. Sometimes when people are thinking, they hold the cup up, right in the way of the eyetracker. Knowing this, one thing we tell users at the start of the session is that "we basically need a clear path between you and the monitor. So if you happen to lean on your hand or hold your coffee cup up, I'll remind you to put it down." That way the user is not surprised, nor does she think the facilitator is too controlling when she gently asks her to put the cup down.

Holding up a cup may impair the eyetracking.

Placing Hand or Finger to Chin or Cheek

Many people put a hand to their chin when they are thinking hard. We usually wait a few seconds before we interrupt the person, as this is often a short-lived gesture. But it does prejudice the tracking.

A hand on the chin may impair the eyetracking.

Leaning on Hand

Conversely, when considering the gesture where the user is all-out leaning on her hand, we find she could endure this for an hour if left alone. So we tend to catch this gesture soon after we notice it as the hand up and the slight leaning adversely affect the tracking. In this position, we typically lose the tracking of at least one of the eyes.

Leaning on a hand may impair the eyetracking.

Hands Up

Some users move their hands to a great extent as they talk, some fidget, and others may pray during a usability session. These gestures are more tolerated by the eyetracker. But if they persist, we do interrupt the user.

Folded hands may impair the eyetracking.

Holding Task Sheet Up

When given a task to read, some users read it and continue to hold up the task sheet as they attempt to do it. We try to give people tasks that are simple to understand in eyetracking studies, so they don't feel they need to refer to it often. You could verbally give users the tasks, which we did for some of the simpler ones. But having the user read the task, we think, helps him to understand it better. Also, he can refer to it later if he needs to without feeling awkward about asking the facilitator to repeat it.

In the eyetracking tests, the paper usually completely obstructed the tracking and so we interrupted the users almost immediately.

If you are doing a think aloud session, asking the user to read the tasks out loud not only enables all of the observers to hear the task that is being attempted, but it also gets the participant used to hearing the sound of his own voice. So, he may be more likely to think aloud than he would if he read the task silently to himself. But, be careful not to embarrass the user with silly tasks. Also, if you think the user may be at all self-conscious about his reading aloud skills, do not ask him to read aloud.

A task sheet may impair the eyetracking.

User Is Writing

In some tasks we ask users to write an essay for an answer. In other cases, some people just like to take notes as they read on the Web. Of course, when the user looks away from the monitor, the tracker loses the eye. We feel it is better to completely lose the eye than it is to get a bad reading. At least when you lose the eye you are not getting bad data.

If the user does write a lot and repositions herself at all, we typically redo the calibration between tasks. Watch the eye status carefully to ensure they are still tracked correctly.

28. Recalibrate the user between tasks if they are writing and shifting in their chair a lot.

User looking away will impair the eyetracking.

Leaning Back

Some users really make themselves at home in the lab, or get tired as they are testing. These people tend to lean way back in the chair. The eyetracker can only keep hold on the eye up to a certain distance (about 60 centimeters, though this also depends on the user's height and angle of the monitor). Even if a user's eye has been perfectly calibrated, if she leans way back in the chair, the eyetracker will lose her.

A user leaning back, getting out of position, may impair the eyetracking.

Leaning Forward

The same principle applies to leaning back and leaning forward. If the user gets too close to the monitor then it will lose the tracking.

We often notice in usability studies that if a website has tiny text, unclear typefaces, or poor contrast between the text and the background that users will lean way in to the screen. This is usually an indicator that one of these text-related issues is present.

29. **If you have a website with tiny text, and unclear typeface, or very poor contrast between the text and the background, try to fix these issues before you spend the resources on doing an eyetracking study. Users will likely lean way in to the monitor when the text is not legible, and their eyetracking will be lost.**

A user leaning too far in may impair the eyetracking.

Squinting

Usually leaning in to the monitor and squinting at it are stances that go hand-in-hand. But even rigid squinting alone can affect the eyetracking.

A user squinting may impair the eyetracking.

Leaning to the Side

When users lean to the side, which is one of the more common arrangements, the tracker will typically lose one eye. This must be corrected.

A user leaning to one side or another may impair the eyetracking.

Combinations

Any of the above postures can, to our chagrin, be combined quite commonly. For example, users lean way over to the side and put their hand up. Sometimes they even squint while doing this. Or they lean way off to the side and slouch far down.

A user leaning with a hand up may impair the eyetracking.

A user slouching and leaning may impair the eyetracking.

All of these positions can be corrected if the user is forewarned at the beginning of the session, and when they are interrupted, it is done gently and not at a pivotal time when she is interacting with the Web.

30. **Warn the user before he begins that you might remind him to move his hand or put down his cup.**

31. **When interrupting the user to move his hand or cup, do it gently, and clearly. For example, "Sorry, Suki. Would you please put the cup on the desk?"**

None of these postures matters in the least for testing without eyetracking. But I (Kara) was conditioned after running so many eyetracking tests. I was doing some non-ET testing for a client and each time a user held up his coffee cup or leaned on his hand, I came dangerously close to interrupting and asking him to put his cup or his hand down. Can you imagine? They would have thought I was a complete spaz**.**

Facilitating Test Sessions

There is plenty to keep track of when facilitating any usability test. In fact, when we present our Three-day Usability Camp seminar, students express the most concerns surrounding keeping things running smoothly when facilitating a test while also observing and understanding what the user is doing. They are far less worried about recruiting users, writing reports, even getting findings from the tests turned into actionable good design ideas. No, they wonder how they will keep so many balls in the air for 90–120 minutes, give or take, when a user is in the lab.

No doubt, the list of things to do is not a short one, from ensuring the user is prepped correctly and is comfortable and on the right track, keeping tasks and questionnaires in order and administering correctly and at the right times, watching any imposed time limits on tasks (to ensure people at least attempt the most important ones in the time allotted), deciding how and when to answer a user's question when it occurs, deciding whether to interrupt the user, and if so how, adding tasks on-the-fly, tracking the time and success and users' final task answers, dealing with the system the user is testing (and more work if this is a prototype or paper prototype), ensuring that observers are getting the most from the sessions, ensuring the video recording and all other test-related equipment is working properly, and last but not least, taking explicit and good notes about what the user is doing. The latter is the point of the test after all: learning about how people approach the design, what's good, what's not good, and most importantly, why.

Even seasoned professionals can sometimes be overwhelmed by all of this. Now drop into this mix a plethora of sequential steps that you need to follow while using a somewhat unstable piece of software and hardware (sorry, but it was, at least when we taxed it with many long tasks), and the need to sometimes say what could be startling things to a user in mid-thought, like, "Would you mind trying not to move the chair back and forth? It really helps if you can keep it relatively still." Odd, right? And hard to act "light" about what you are saying when you feel besieged with administrate trivia.

All of this will unnerve even the most hardened usability professional. Rattled as we might be, we can all do these tests and do them well. And here are our suggestions for a few more things to ensure they run smoothly.

STEPS FOR MAKING EYETRACKING TEST FACILITATION RUN SMOOTHLY

32. **Disclose the basic eyetracking-related information in recruiting, but don't overdo it and make the user nervous. We already discussed the importance of starting the session right and explaining things to the user and calibrating them well. These, and even the brief explanation in the phone recruiting, are the cornerstones to a successful eyetracking test session.**

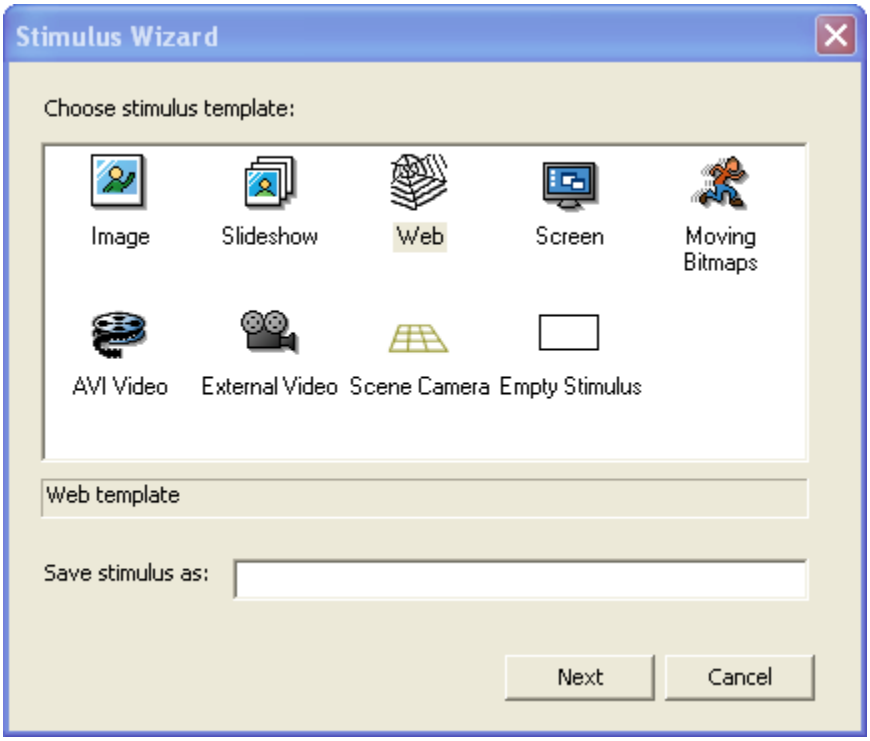
33. **Recruit the right people for maximum eyetracking calibration.**

34. **Prepare all tasks in the eyetracking software before the user arrives.**

35. **Get done whatever you can with the eyetracking technology before the user arrives. For example, in the software we used, each task is**

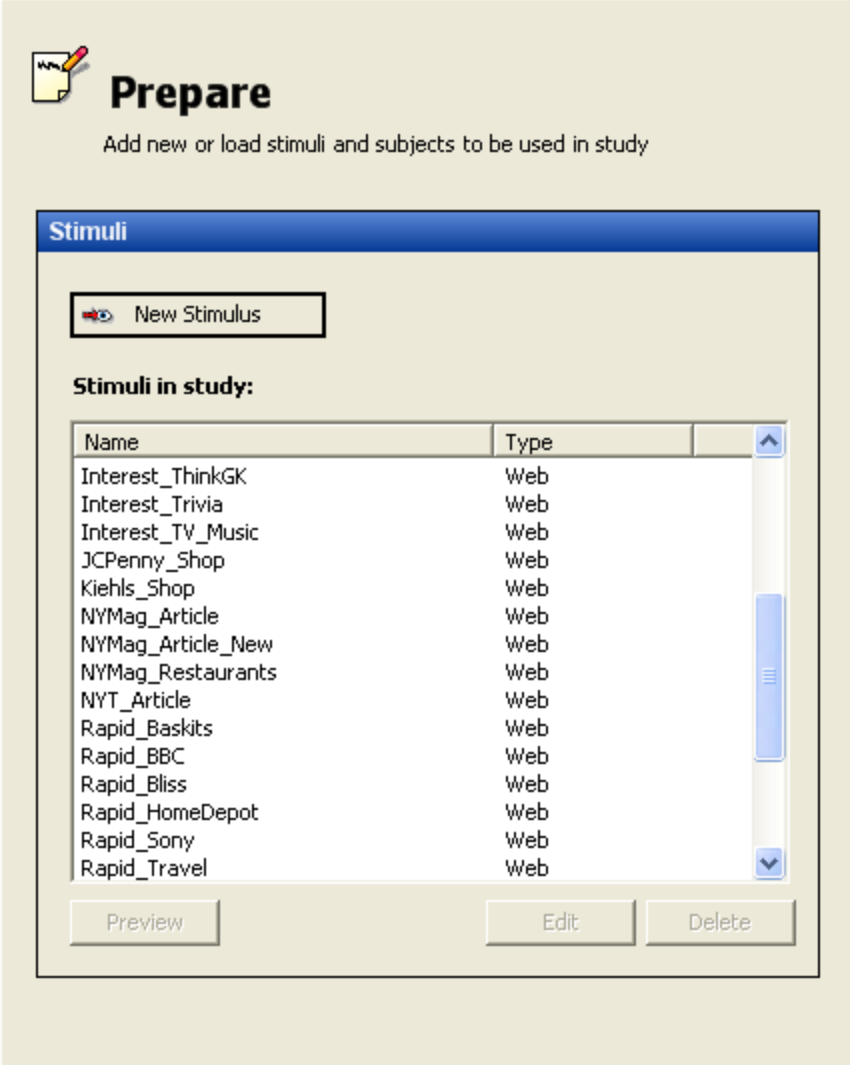
saved as a separate "stimulus" in the system. So when you ask the user to "find the headquarters for the Agere company," you, before the test, create a stimulus and define any parameters in it, such as blocking pop-ups or telling the page to load incrementally or all at once.

We typically used the "Web" stimulus template in the eyetracking software because it launched a Webpage for us. Even if we didn't want to direct the user to a particular site to test, we might still use a Web stimulus to launch a browser and start them on, say, some newspaper website. All were automatically launched when we started the stimulus. From there the users could go to the search engine or type the URL of their choice.



You may choose the type of stimulus to record.

36. Make the names of the tasks you ask users to do easy to see in the list of tasks you are using in the ET software.



This section of the dialog box shows the list of stimuli, or tasks, created before the users arrive onsite. In the Clearview software UI that controls the eyetracker we used, the facilitator chooses a stimulus for each task just before the user will attempt to do it. Selecting the right stimulus will launch the website you assigned, start the eyetracking, and record the session until the facilitator stops it. When testing many different tasks, it's helpful to give them distinctive names so the facilitator can easily distinguish between them at a glance between test tasks.

37. Since Webpages typically do load incrementally, for most tasks direct the eyetracking software to also load them incrementally.

Since Webpages typically do load incrementally, for most of our tasks we had the eyetracking software also load them incrementally. (Following one of the ground rules for all usability: to be as representative of real use as possible, so that the test findings will be applicable to improving the design for real-world use.) The one exception occurred in our "rapid fire" tasks when we wanted to know where users look first for a specific widget in the UI. For these tasks, we had the eyetracking software load the whole page at once, not in pieces.

Also, for this task we wanted to launch the site for the user.

38. **Test/do all tasks with the eyetracker running to ensure there are no glitches (ideally, pilot-test all tasks with a colleague).**

39. **Check the recordings from the tests to ensure you are getting the data you need. (This is where you may learn there are too many dynamic elements to be able to use any heatmaps accurately, for example.)**

40. **After doing between-session cleanup (such as clearing the browser's cache), if using a dual-screen set-up (so the facilitator can watch on a separate monitor) ensure that the browser window appears on user's PC, not facilitator's, and is maximized when you launch the stimulus in the eyetracking program.**

41. **Set up your facilitation workspace so you can easily get to the user's task list pile, the post-task questionnaire, your stopwatch, etc. This is important in all usability tests but especially important in ET tests because, 1) you have so much more to do and need to keep the steps straight, and 2) when you hand the user his task sheet, you need to wheel over to him and do it right when he is ready for the next task before he begins to move or fidget in his seat.**

With all the things to keep track of when facilitating an eyetracking session, the last thing you want to do is drop your stopwatch, thus potentially losing the task timing and possibly freaking out the test participant. Use an online tool if you prefer, or just tie the string around the facilitator's chair on the side away from the participant, and you will always be able to quickly locate your trusty timer.

42. **Tie your stopwatch's string around the facilitator's chair on the side away from the participant so you can always quickly locate your trusty timer.**

The facilitator uses the stopwatch's string to attach it to the chair, ensuring it doesn't drop or get lost during the study.

43. Have a cheat sheet handy with the user's name and file naming conventions. So when you are doing 50 things at once and the *Save* dialog appears, you can quickly type the right name that will help you find and decipher the files later. We just stuck a Post-it-note on our facilitation monitor with the file naming scheme on it.

Rather than having a Post-it on your facilitation monitor, you could keep one right on the PC's clipboard. But we were doing a dual-monitor setup, so the facilitator and user shared the same clipboard. If the user had used the Windows clipboards and

tried to paste something, albeit a rare thing to do in our Web studies, he would have pasted something like “ex_User2_Mike_H.”

44. **Create a test tasks and protocol sheet that has everything you need to say and do in sequential order. Over time, if you do many tests, you may not need this anymore. We didn’t by our 150th user. ;)**

45. **When the user arrives, prepare and set up thoroughly, and make him comfortable in the lab.**

46. **Take the time to calibrate the user right.**

47. **Keep the test session moving even while you are waiting for the eyetracking software to save a file or grab a file. For example, when telling the user to save the file at the end of a task, while waiting for the system to open the *Save* dialog box, give the user the post-task questionnaire, turn to your next list of things to do in your protocol sheet, and get the next task sheet ready to hand to the user.**

48. **Backup all files on an external hard-drive at the end of each testing day, for safety reasons. Set the program to run overnight, as exporting ET files often takes several hours.**

EXAMPLE OF FACILITATOR'S PROTOCOL SHEET

Round One								
Task	Stimuli	Task Goal	Tasks	Points the User should mention on Task Repeat / Facilitator confirm	Task Answers / Points	Where to Go in The UI	URL
1	**Agere_HQ**	BROWSE / FIND Co. HQ	Please use the **Agere** website to find out where the company's headquarters is located. Please remember to tell me when you are finished and to tell me your answer once you have it.	** Find Headquarters	Allentown, PA-4; OR PA-3	> About Agere > Fast Facts > Scroll	agere.com
		ASK:	Would you please summarize what that is asking of you?				
		SAY:	Once you are ready, please look at the screen. I will open the website for you, and then you can do whatever you'd like.				
		DO:	Look for the two eyes in ClearView.				
		DO:	Click the *Start* button in CV.				
		SAY:	(If there is a time lag) The site will load in a moment.				
		DO:	When site is loading and the user is addressing the system, start the stopwatch.				
		USER:	User does the scenario.				
		SAY:	(When user gives answer and has indicated he has finished) Say: I am just going to take control of your mouse for a moment, so if you would please let go of it.				
		DO:	Wait for them to let go of mouse, then stop Tobii recording (press F10).				
		DO:	Stop the timer. Note task time, user's answer, and if possible rate the success score (May also do later).				
		SAY:	(While waiting for CV to get the *Save* dialog up . . .) Please answer this questionnaire -- these four questions. (Gesture to the four questions.)				
		USER:	User answers the post-scenario questionnaire.				
		DO:	When the *Save* dialog appears in CV, save the file with appropriate info, such as "ex," "U1," "Mary," "F," "H"				
		DO:	Select the next stimulus in ClearView. Click *Record* button. CV will take a moment to launch it. Wait for it.				
		DO:	Check the user's answers in his questionnaire to ensure all answers have been given and the user is interpreting the numeric scale correctly.				
		DO:	If the calibration was dropping in the last task, recalibrate now.				
		DO:	Get the user's next task sheet from the pile.				
		SAY:	Ready for the next scenario?				
		DO:	Hand the user the scenario sheet.				
	Next Stimulus	**Adelphia_Browse**					

SAVING FILES

One way you can make analyzing data from the sessions easier is to put pertinent information right in the filename for the tasks as you save them real-time.

Since we were running large studies with hundreds of users, each user was given a participant number. This made dealing with the background and syncing it with the ET files a lot easier. But to accommodate possible typos or duplicates, we also included the user's first name in the filename.[5] For example, we'd save the file of our first participant, who happened to be named Mary, as *U1Mary.*

49. To make analyzing data from the sessions easier, put pertinent information right in the user's task filename for the tasks as you save them real-time.

We also thought one of the most important things to note was the quality of the eyetracking. If the tracker was good for the file then we did not comment on this at all. If the user's eye was not tracked or poorly tracked, we used the abbreviation "ex" to mean to exclude. We didn't want to write the whole word *exclude* because if the user somehow got a glimpse, we didn't want them to feel as though something was wrong. Also, *exclude* takes a lot longer to write than *ex* does. So if Mary was poorly tracked, we'd save her file as: *ex_U1Mary.* With this filename, we know to deselect this item in any heatmaps we draw, and not to bother creating gazeplots for this user.

If the user's eye seemed to get lost and was in and out sometimes, but we were not sure if the track was actually bad, we typed *med,* short for *medium,* to mean we needed to thoroughly watch the gaze replay before we used any of the file for gazeplots or in heatmaps. So if Mary was mostly tracked well, but there were a few questionable spots, we'd name the file: *medU1Mary.*

Another item we liked to save in the filenames included Web experience. We delineated these with *H* in the filename for *high*, *M* for *medium,* and *L* for *low.*

50. In the user's task filename, indicate whether there was poor calibration and to what degree, so you can later decide whether to use any of the data.

When interviewing the user after some tasks, we wanted to record the interview, so we used the eyetracker because it was set up and easy to do so. But we did not want these interview sessions to get mixed up in the pure test sessions. So we'd always add the words *Interview* at the beginning of any interview-type tasks. So when choosing users from the list to add to the heatmap, we never selected those with *Interview* preceding their user number and name. (As noted earlier in this report, users' eyes roam very differently over a page during a free-form conversation than they do during realistic use. Since you're designing websites to be used, not to be discussion pieces, you want ET data from use, not interviews.)

[5] For privacy reasons, you should avoid associating the user's full name with the recordings of their behavior. In fact, only write the user's full name in those few places where it's absolutely necessary, such as the consent form and the receipt for the test incentive. (In case the tax people ever audit you for having distributed $25,000 in unmarked $20 bills, you want to have the documentation in perfect order for having given $100 to each of 250 test participants.)

51. In the user's task filename, indicate whether the file is an interview, not a task.

We also found it helpful to note gender. Writing an *F* for *female* in the filename does the trick. No need to write an M for male because not only is it not necessary, but it could get confusing with the M for medium Web experience. Thusly, if we wanted to spot-check heatmaps comparing men and women, we could easily scan the list of users who attempted a task and choose all the *F* files to include only the females in the heatmap.

52. In the user's task filename, indicate any background information you think is important, such as gender or Web experience level.

If the user has made the trip to the lab but his eye cannot be tracked at all, consider running the test anyway without eyetracking but adding the think aloud protocol. This can give some richness to the eyetracking data you get from the other users. Make lemonade out of lemons.

RETROSPECTIVE

Usability test facilitators have long weighed the pros and cons of interviewing users after usability tests.[6] The process called *retrospective* is a different twist on the typical after-the-test interview. This process has a few more benefits but, unfortunately, still has the same drawback. With retrospective you show the user the videotape of himself during the usability test as he used the design. In playing the tape for the user, you can ask him to elaborate on what he was doing and thinking during the testing. The video is meant to relieve some of the memory pressures on the user and in fact, jog his memory. And asking him the questions while watching the user's actions "real-time" can direct the user to the exact areas you are interested in and that he seemed interested in or confused by. Of course, there are benefits and drawbacks to using this method.

On the one hand, you can get questions answered by asking people. On the other hand, users may fabricate answers and nullify all the good behavioral research you just spent more than an hour collecting. They can post-rationalize, feeling they need to defend some of the actions they took rather than just discussing them. Also, it can be very time-consuming to fast forward, rewind, find, play, then ask about the actions the user took. And directing a user to a particular point in his videotape may force him to think a little harder about what he was doing at that point. He may think what he is saying aloud in the retrospective is, indeed, what he was thinking during the test when it is really not.

Retrospective emerges as a methodology in eyetracking research in particular because of the issues that think aloud protocol imposes. Instead of think aloud, which is not often used, try the retrospective method to get some user discussion and quotes.

---

[6] Jakob's dad, Gerhard Nielsen, conducted research in 1956–57 on the topic of asking study participants to comment retrospectively on movie recordings of their behavior. It's not a new idea. (In the 1950s, the older Dr. Nielsen used a film camera that could only record 6 minutes of behavior at a time, and even this primitive technology was not available at Copenhagen University, so he did the research at Harvard University, benefiting from a Fulbright exchange fellowship. So while the idea is not new, the execution has certainly become easier, with webcams, Morae, and the like.)

We say stop trying so hard to get people to verbalize and concentrate on analyzing behavior.

Also, the obvious downside of any replay-based methodology is that running the replay consumes extra time when you're not observing new behaviors but simply revisiting ones you've already seen. In the time it takes to have a user provide retrospective (and possibly rationalized) comments, you could either have tested another person or had the same user attempt additional tasks and/or tested another design variant. We usually prefer breadth over depth, because the design space is so vast in interaction design that there's always more to explore. (Whereas the issues are usually simple enough that a trained usability expert can identify them without too much painstaking digging into the minutiae.)

53. Consider carefully whether retrospective would actually help, or whether you can get the information from observation. We typically do not recommend using retrospective in ET.

In doing retrospective in eyetracking research there are the above considerations to account for, and more. For example, when showing the user his eyetracking file, even if you explained to him that "the blue dot represents the way his eye moved," that "all eyes move around fast and jerkily like you will see" people are still shocked, sometimes nervous, and at the same time, enamored with the gaze replay. If we had a nickel for every time the following conversation, or one like it, played out in our tests:

Play gaze replay

User: Is that my eye?

Facilitator: Yes, it is. As I mentioned, the blue dot represents when you looked at something. As it gets bigger is means you looked at the object longer.

User: Oh.

Facilitator: Let's take a look at this page. Can you tell me a little about what you were thinking here?

User: That's my eye?

Facilitator: Yes, it is. Now about the page . . .

User: I can't believe that's my eye.

Facilitator: Silent waiting

User: Is that normal, the way it is moving all around like that?

Facilitator: Yes, it is. There's nothing to worry about because everyone's eye moves like that. Pause. So, about this screen . . .

User: Oh, yes. So, I am looking at that button right then? That's my eye?

> Facilitator: Please pardon me for a moment (while I go shoot myself in the head.) You keep watching the gaze replay.

Of course, we had this same kind of reaction ourselves when we first tried using eyetracking—we were utterly awe-struck and probably also unable to concentrate on any retrospective. But even beyond the vanity-related issues associated with watching your own eye, we see other issues with using retrospective in eyetracking, and in general, we do not recommend employing this method. As mentioned previously, there is too great of a chance that users will accidentally fabricate thoughts about what they were thinking as they looked at items on Webpages. Half the time we as Web users are moving too fast to even have fully-fabricated thoughts, yet in retrospective we are asking for articulate commentary.

If you do decide to use retrospective, ensure that you explain to the user, well before you launch his file, that rapid eye movement is completely normal. Explain how the blue dot movement and fixation length works, so they understand that the dot held on a point means they looked longer. Also tell them about saccades, represented by lines that show up between dots, and that these lines do not indicate that the user looked at the items under them. (In fact, a person is basically blind during a saccade when fixation moves from one point to another. But don't tell the user that.) Only a dot indicates a look.

Alternatively, show the replay but do not show the user his gaze at all. But doing this can make the retrospective even harder to follow, especially since the user was not thinking aloud during the original recording. (Assuming that you want to record "pure" eyetracking and save the user's comments for the retrospective.)

At all cost, do not show the user any of his videotape with or without gaze replays until the entire testing part of the session has been completed. When we pilot-tested doing this between tasks, it really freaked people out or they were really fascinated with it. It made them think about what they were doing with their eyes as they were attempting their subsequent tasks, which was not at all good.

54. If you do decide to employ retrospective, explain to the user that the blue dot moves quickly and there is nothing wrong with his eye. Alternatively, show only his video without the gaze replay.

55. If you do decide to employ retrospective, always conduct this interview at the very end of the session. Do not attempt to do any behavioral research once this is done.

Robert Stevens, co-founder of Bunnyfoot, a British usability firm, is founder and advocate for a post-eyetracking-test interview method called PEEP (Post Experience Eyetracking Protocol). PEEP is based on retrospective but has some variations, the main one being that the retrospective interview is an augmented by insights the moderator learned when watching the real time tasks.

He explains, "The eyetracker allows the moderator to peep into the real time subconscious thought processes occurring in the user, you can't get that with just a screen recording."

Rob also notes that users are too entranced with their own eye to actually warrant showing them their actual gaze reply, and now opts to show the video replay with the gaze tuned off. He explains, "The PEEP methodology has evolved since we first developed it in 2003. We used to show the user parts of their user journey to prompt their memory. But this is often too engaging for the user as they have usually never seen an eye tracker before. They find the technology amazing and concentrate on

the eyetracking rather than on their actual journey. In recent years we have moved away from showing the user their gaze replays to talking about it."

Using the Right Eyetracking Analysis Tools

We discussed heatmaps earlier in this report, but there are three very good analysis tools in the eyetracking technology that we worked with in our largest ET study: gaze replays, gazeplots, and heatmaps. This section discusses more specifics about using all of these ET representations. They each have their good points and bad points, and we find that using a combination of all three is important to getting valid and accurate information.

GAZE REPLAYS

Gaze replays enable you to watch the test session video again with the user's eye motion overlaid over a recording of the changing computer screen image. In these you can see a blue dot, or user's eye, moving around a page, and hear the user's voice. In some cases you may also record the user's face with a PC video camera, such as a webcam. (But we had to forgo this video as our sessions and each task tended to be very long and thus the video was too taxing and was crashing the eyetracking system.)

Gaze replays are the most valuable, in our opinion, because you can slow down the replay and really see everything the user looked at in the order in which they were doing it. (Of course, for the think aloud sessions all you hear when the video is slowed is "waah waaaah waaaaaaaah," so the audio is not valuable unless you watch the sessions at full speed.)

Gaze replays are also the most accurate method for analyzing the information. We later discuss heatmaps and gazeplots, also good visual representations but ones that have faults. Namely, with heatmaps you get no sense of the order in which people looked at items. With gazeplots you get some sense of this, but it is sometimes hard to see. More importantly, we encountered some gazeplot displays that dropped the session part-way through. So the gazeplot showed a user stopped looking at the page at a certain point, but really there were many more fixations than the plot showed in the actual gaze replay. Most cases where this happened, the lock on the user's eye wasn't completely strong.

Gaze Replays are Time-Consuming to Watch

The drawback with using the gaze replay analysis is that it takes a long time to watch the videos even just at regular speed. And if you slow them down they take even longer. (We usually find that watching at one-third speed is best for understanding most of the details of a user's eye movements, and that's the replay speed we use for the slow-motion segments we show in our seminars.) The eye moves so quickly you often do have to slow and replay segments to get an accurate idea of what happened. So if you have interested employees who have a good segment of time to dedicate to watching these replays, then eyetracking may be for you. If not, we recommend you reconsider doing eyetracking studies.

When we looked at our notes we took during the session we had comments about fixations that did not appear in the gazeplot. We first thought our notes must have coincided with a different user or something. But when we watched the user's gaze replay, it revealed that he did, indeed, look at many more things than the gazeplot depicted. This is not typically a problem with the gazeplots, for sure. But it happened often enough that we needed to check for this mishap in order to be confident in the gazeplots we were using.

Gaze Replays are Interesting

If you are thinking that people will get bored watching the eyetracking gaze replays, you may be as surprised as we were. Watching the gaze replays is far more interesting than watching typical usability test session videos after-the-fact. And watching the sessions live is downright riveting. Even observers who are not usability geeks were quickly engrossed in these sessions. It gives most people the feeling that they are actually getting into the user's mind as he works with the interface.

GAZEPLOTS

Gazeplots are another very valuable analysis tool in eyetracking technology. These compile the eye gaze of one user on one page. These do not combine more than one page or user in the representation. Light blue dots on the page show a number of important data items, including: 1) where the user's fixations were, 2) numbers in the dots depicting the order in which the user looked at the items, and 3) the size of the dot denoting how long the user looked at the item. Larger dots mean longer looks. Also, you may use a timing slider to depict parts of a task session. So if there are a zillion dots on a page but you just want to see the dots related to the first 10 seconds on the page, you can look at that, which is helpful.

Sometimes Dots on Gazeplots are Difficult to Differentiate

The limitations of gazeplots all have to do with the visual representation of the dots on the page. It is very hard or impossible to see the numbers on the dots when there is more than one in an area. Also, the dots sometimes impair the page itself. So you see the user's dots in an area and know he is interested in that area, but the dots themselves are covering the page so you can't see what the user is interested in. (Oh, the irony—these dots are telling me to pay attention to the stuff under them, but the very dots are not allowing me to do so.) You can adjust the transparencies of the dots and view the page with no dots at all, which is helpful.

Gazeplots Can be Time-Consuming to Analyze

Another limitation of gazeplots is that, again, they are very time-consuming to work with. For example, say 25 users looked at the same five screens and you want to see their gazeplots. Of course, you could make a heatmap instead of a gazeplot, but a gazeplot is better if different users spent very different amounts of time on a page, because it can be misleading to average people with very different behaviors. Also, the heatmaps give no indication of the order in which people looked at the items—you just know they did. (See the heatmaps summary for more info on this.)

In gazeplots you see only one user at a time. So to look at the five pages, you have to create gazeplots for one user, find the screen you are interested in from a list of sometimes many, and then capture it. Then find the second screen in the list and capture it. Then do this for the remaining three screens. One user down. Then you must repeat this whole process for the other 24 people. This can easily take an hour or more to do, and that doesn't include any analysis at all—just grabbing the screens to look at.

Issues with Dynamic Elements

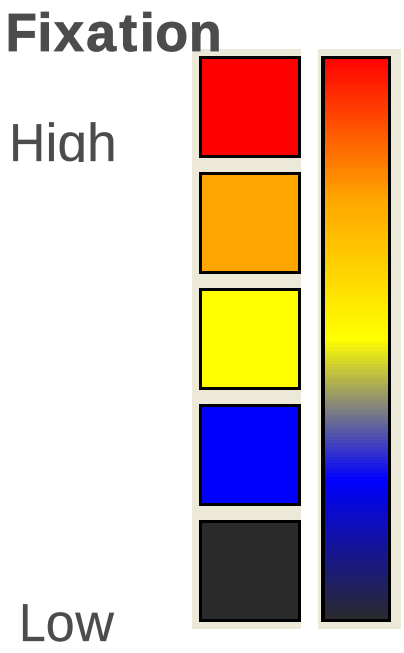
Finally, an important and problematic limitation of the gazeplots, which we also see in the heatmap analysis method, is that gazeplots do not show drop-down menus, pop-up windows, video, or other dynamic Web elements. (Newer versions of the ET software may address this.) And there is no indication when the gazeplot is wrong.

The gazeplot just shows the dots as though the user were looking at the items on the page, with no indication that they were actually looking at an item that appeared over the page. For this reason, it is imperative that you do not just create gazeplots and assume they are accurate or that they tell the full story. The following section about heatmaps will cover more on this topic.

HEATMAPS

Heatmaps visually display the areas where users looked on a Webpage. You may create a heatmap with as few as one user or with many users. You can choose to create heatmaps showing the fixation length, in other words, time, or number of fixations. We chose to create our heatmaps showing the fixation time instead of the number of fixations because we felt that one long fixation should count for more than a short one. If we made heatmaps showing number of fixations then we would miss that, as all fixations, long or short, are simply counted equally. Either depiction would be telling though. In the sample heatmaps we ran, we did not see a tremendous difference in them when we created them based on length or number of fixations.

In our heatmaps we decided that the most heat should be red, then orange, then yellow. As areas got fewer looks they are represented with blue and finally, a dark, black but still transparent look.



You may choose the colors to indicate more or less fixation in a heatmap. This is the scheme we use.

In most eyetracking research reports of late, including our own, we see an inordinate amount of heatmaps. These are great for getting a quick overview of how users looked at a particular page. And they are a very good start in exploring where to further analyze a page. They must be taken as part of an entire process though. Appraising a heatmap in a vacuum can be the root of all sorts of misreads. There are a few very common reasons for misreads in heatmaps. They are:

- The **user is talking** about something or thinking aloud.
- There are **dynamic elements** on the page. (This issue may be fixed in newer versions of the technology.)
- **A page is being evaluated alone, not as part of the sinuous Web process, and does not consider *user recall.***
- There is **no depiction of order** in which items were looked at.

Color Scheme for Heatmaps

The color-coding in heatmaps is completely arbitrary. There's nothing that says that the most-viewed parts of the screen should be red and that the least-viewed parts should be blue. However, this color scheme is used by many eyetracking researchers and is the reason for the very word "heatmap." (It's called a heatmap because red is hot and blue is cold.)

We experimented with several other color schemes before deciding on the red-yellow-blue scheme, which we also recommend that you use. The examples here show some of our ideas.

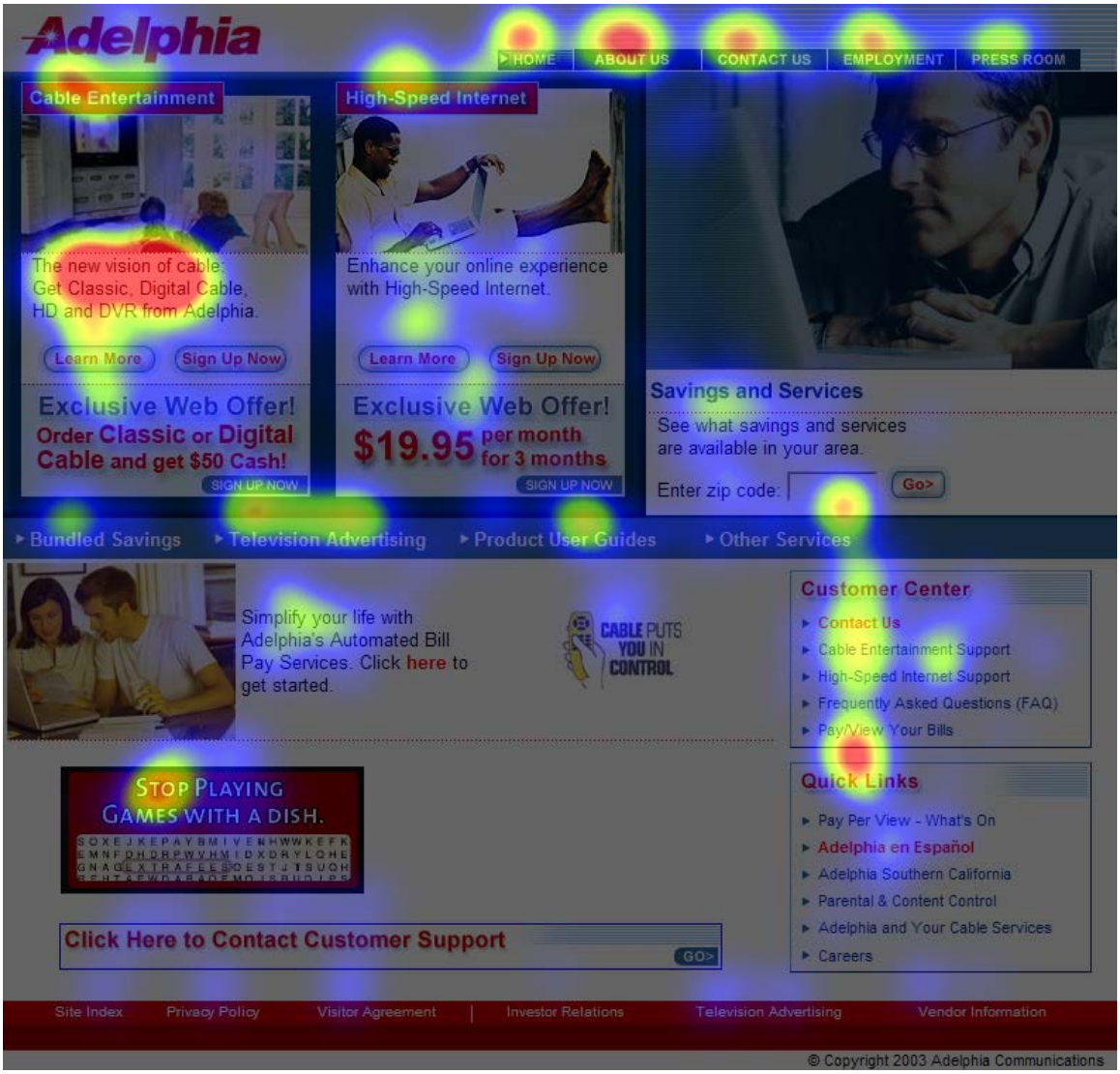
The grayscale coding is probably the most scientifically correct, because there's actually a true progression between the different shades of gray that reflects the ordering of the underlying data. Thus, there is no need to explain the coding and no need for the researcher to learn the meaning of the colors.

Despite these analytic advantages, we ended up not using the grayscale coloring because it's less dramatic than the red-yellow-blue scheme. When we show eyetracking slides in our seminars, we want the audience to pay attention. And, of course, we also want our own ET research findings to be colorful and attractive to readers.

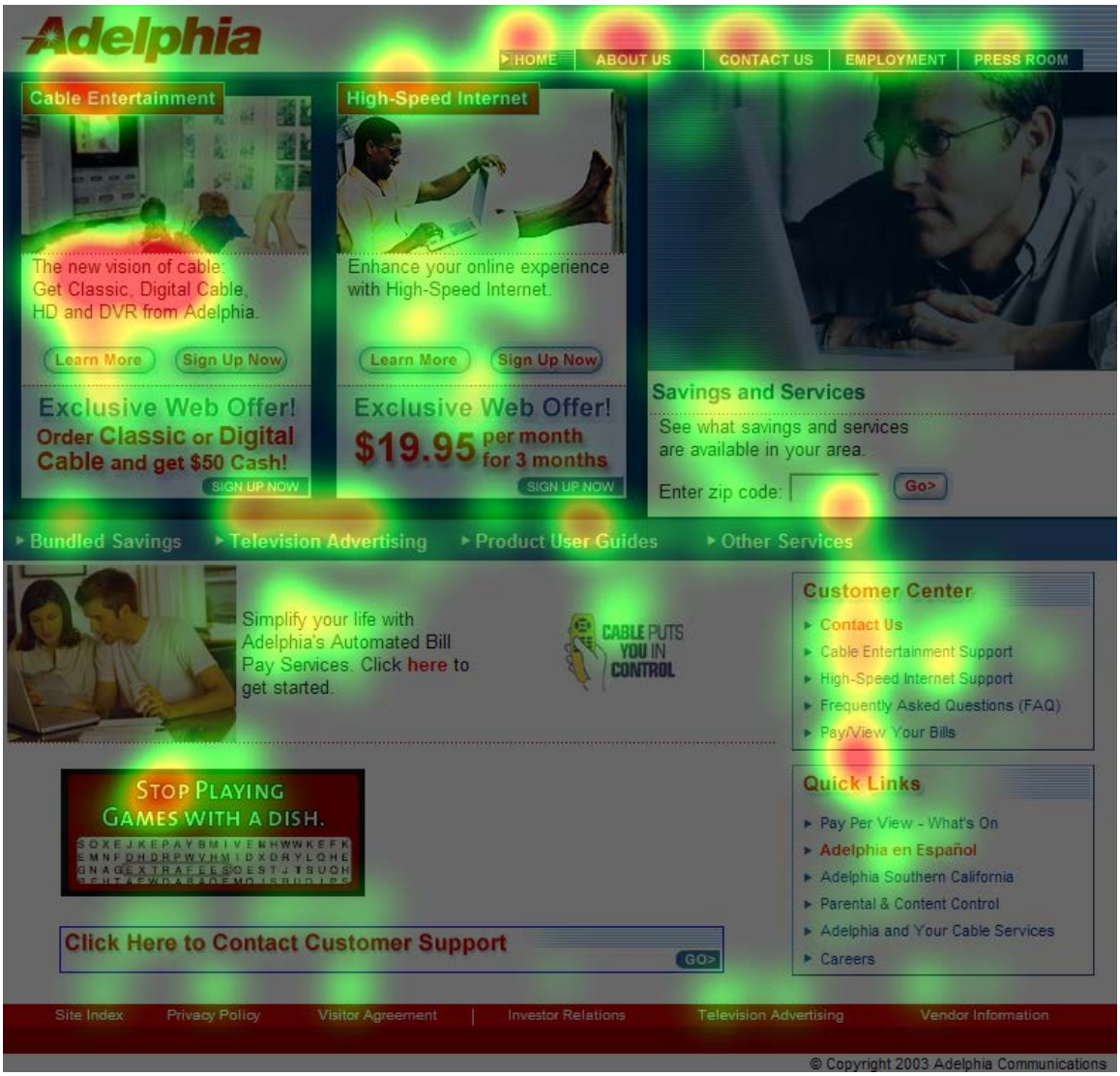
Finally, there's an advantage to using the same color scheme as most other eyetracking reports.

We decided to make the background dark gray as opposed to the white we also experimented with (and which is shown in the last illustration in this section). Those parts of the Web page that users never looked at should recede into the background, which is achieved by the dark color. In contrast, the white background color makes the unviewed parts of the page the most readable in the heatmap and thus presents a distorted impression of the users' behavior.

The following examples shows different color schemes used to encode exactly the same data, so that you can judge for yourself which approach is the best visualization.



The color scheme we chose for heatmaps: the most-viewed areas are red, less-viewed areas are orange or yellow, the least-viewed areas are blue, and areas that didn't get any fixations are gray. This is the homepage of www.adelphia.com.



An alternative color scheme that was actually the default option in the eyetracking software we used: the most-viewed areas are still red, but the least-viewed areas are green, which is a very bright color and gives these areas an undue visual prominence. This is the homepage of www.adelphia.com.

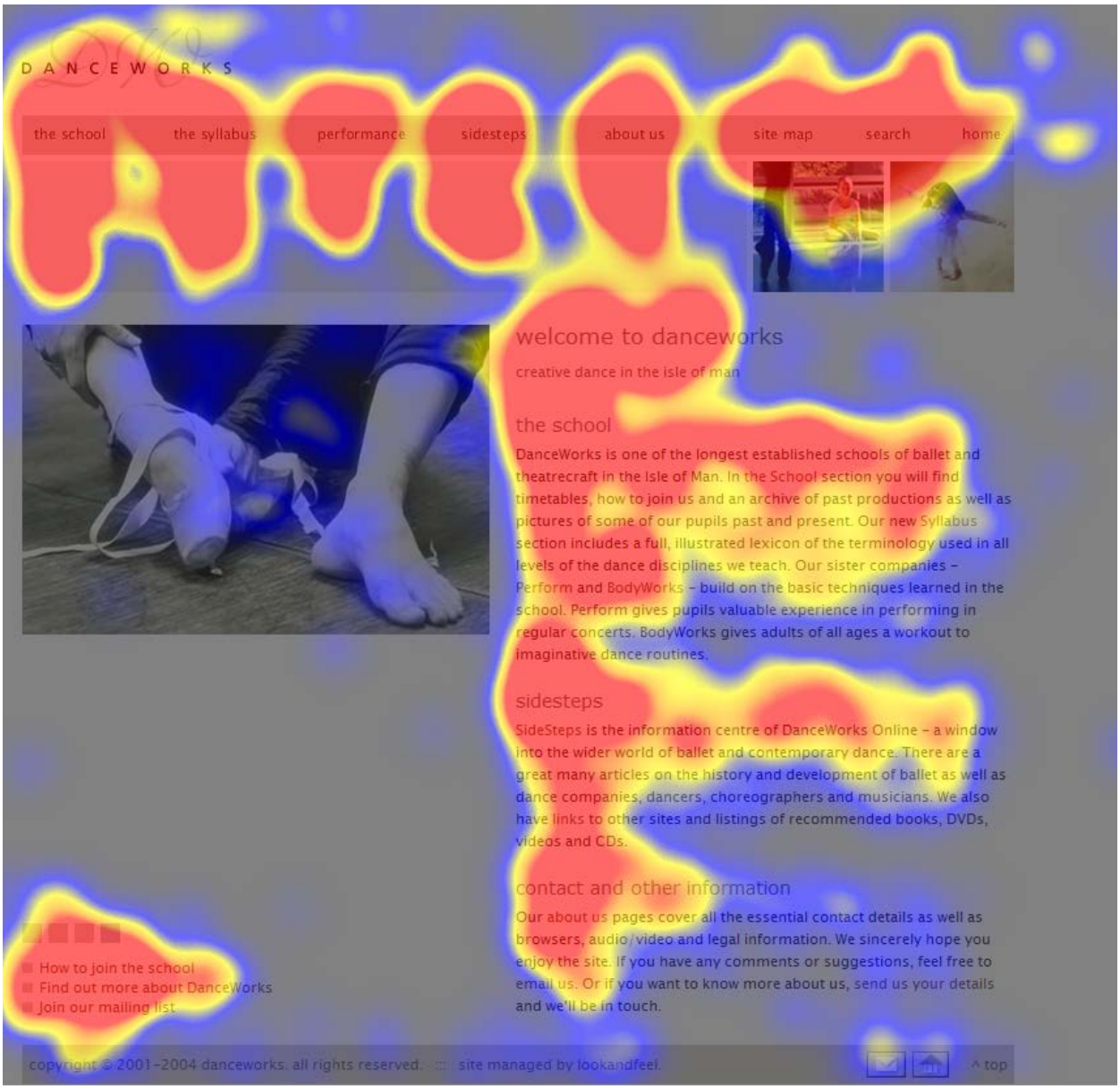
A grayscale color scheme, where the most-viewed areas are white and less-viewed areas are progressively darker. This is the homepage of www.adelphia.com.

An alternative color scheme (or a different page this time) where areas that were not viewed are coded white. On the one hand, this scheme makes it very clear where users looked, but on the other hand, it gives undue prominence to the little-viewed blue areas, which have a very high contrast with the white background. In fact, on the white background, the yellow medium-viewed areas are less prominent than the blue areas, even though the opposite would be desired. This is a page on www.akc.org.

Dynamic Web Elements in Heatmaps

Evaluating pages that had any sort of dynamic elements, such as multiple windows, pop-ups, drop-downs, or cascading menus caused misrepresentations in the heatmaps—at least with the version of the technology and accompanying software that we tested with. Please consult ET technology websites for updates and fixes to these issues.

When watching the sessions live, you know if the user is looking at a window that popped over the Webpage they are on, and not the element on the Webpage that is in the same spot under that item. For example, a user may read the items in a drop-down menu, but when you analyze the page, via a heatmap or gazeplot of the user's look on the page, you do not see the drop-down menus. Instead, all you see is fixation activity where the item was, but the heatmap created with the eyetracking technology does not display the menus. An example of this occurred on the DanceWorks website. Users were looking at the menus as they opened in the top horizontal menu bar. Watching live (or when reviewing the replay in the eyetracker) you can see the menus dropping down. The users looked at the menu items, but in the heatmap it looks like users were just looking at the white space below the menu.



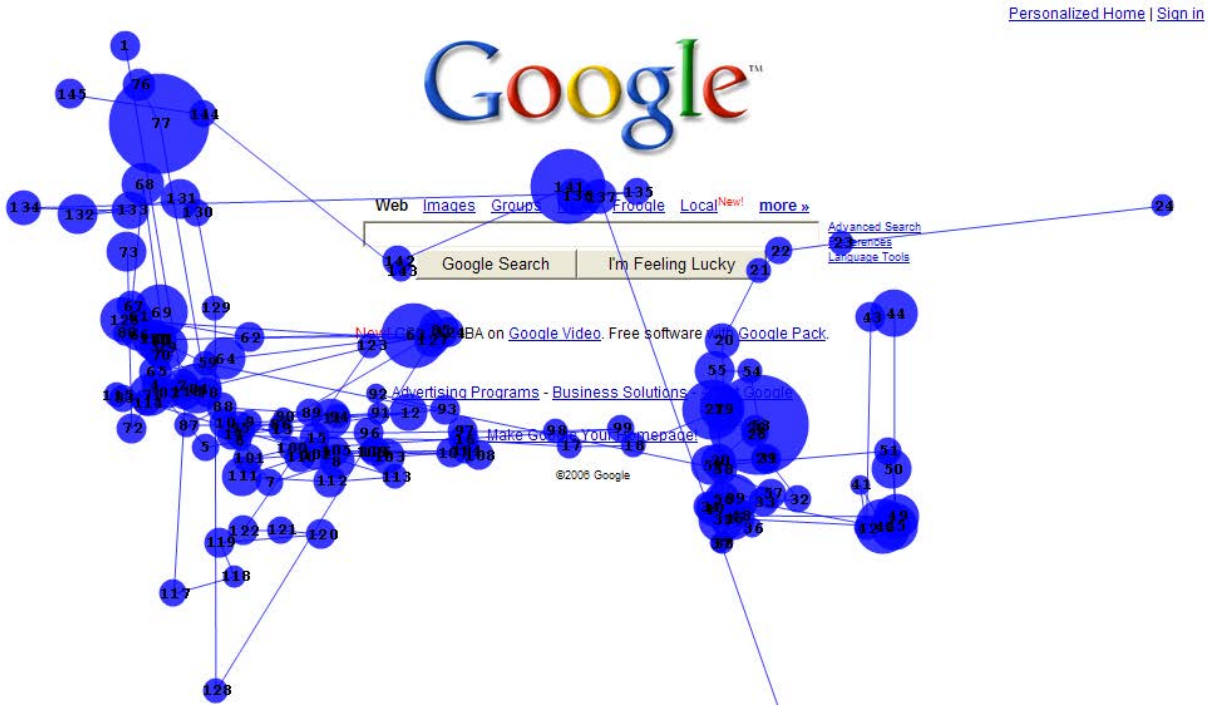
Misrepresentations are heatmaps or gazeplots created with the eyetracking technology that depict heat or gazeplots on a page inaccurately because of a limitation in the technology. Namely, the user is looking at an item that temporarily appeared real-time over the page, such as a drop-down menu, a pop-up window, or another browser window, that is not actually depicted in the static screenshot underlying the heatmap. Only sometimes are misrepresentations obvious. You must watch the gaze replay (or live session) to truly know.

56. Look out for *misrepresentations* in heatmaps and gazeplots.

We call these inaccuracies in heatmaps and gazeplots *misrepresentations*. Misrepresentations display in heatmaps or gazeplots created with the eyetracking technology. Heat or dots are depicted on a page inaccurately because of a limitation in the technology. Namely, the user looked at a dynamic element that appeared real-time over the page, such as a drop-down menu, a pop-up window, or another

browser window. But this item is never reflected the heatmap. So it looks as though users were staring at some element on the page (or if you are lucky just white space on the page) but they actually were not looking at that at all.

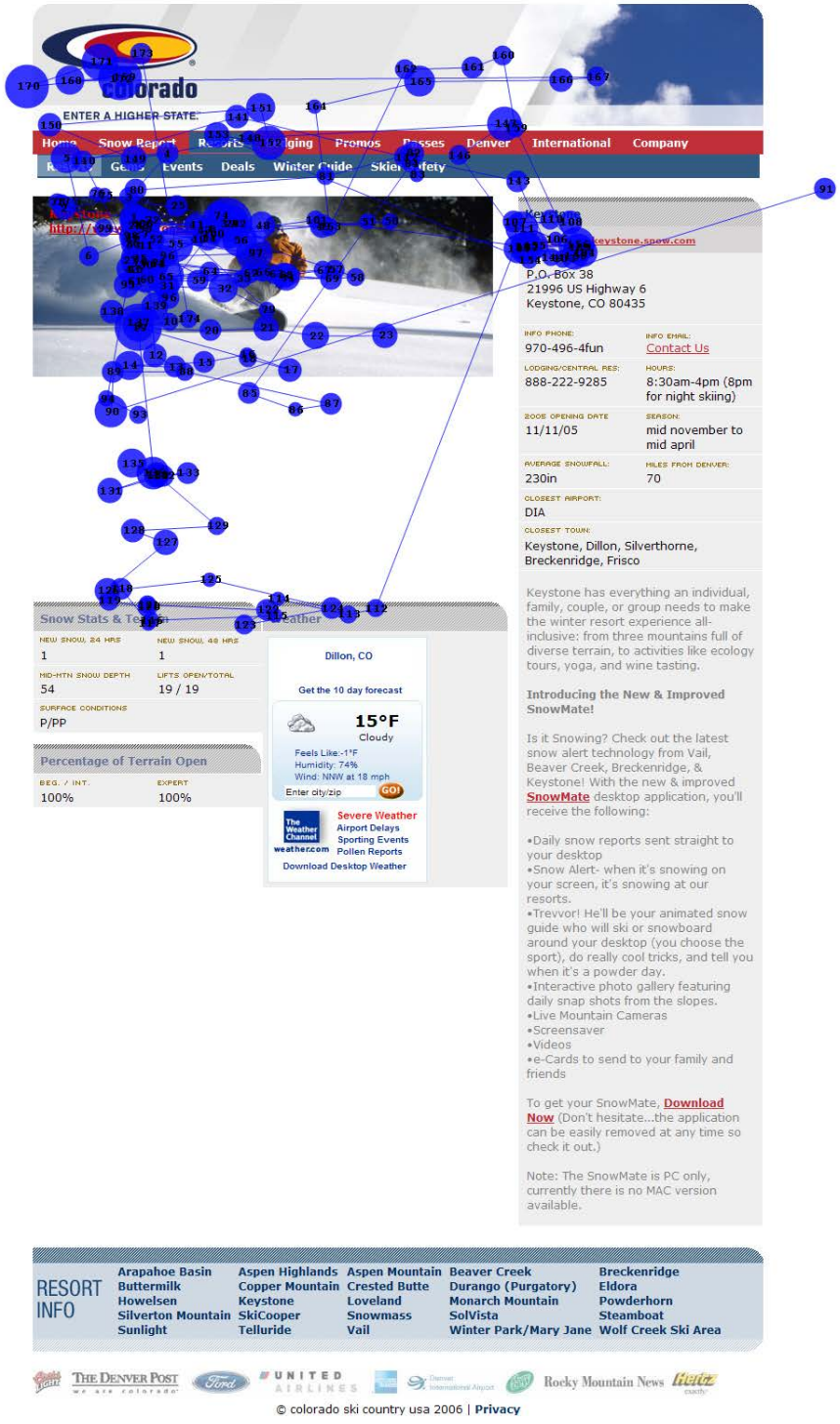
Gazeplots in places where there is no content, like the following Google page, designates something is wrong.



An example of a misrepresentation on a gazeplot on www.google.com.

Heavy heat on a white part of a page is an indication that there might be a misrepresentation in a heatmap. Cascading menu misrepresentations are also among the easiest misrepresentations to spot in heatmaps, as there is heat just below the menu items that are, indeed, visible in the heatmap. (Heat pattern looks like the image above.) Random looks on a page are also an indication that there may be some misrepresentations.

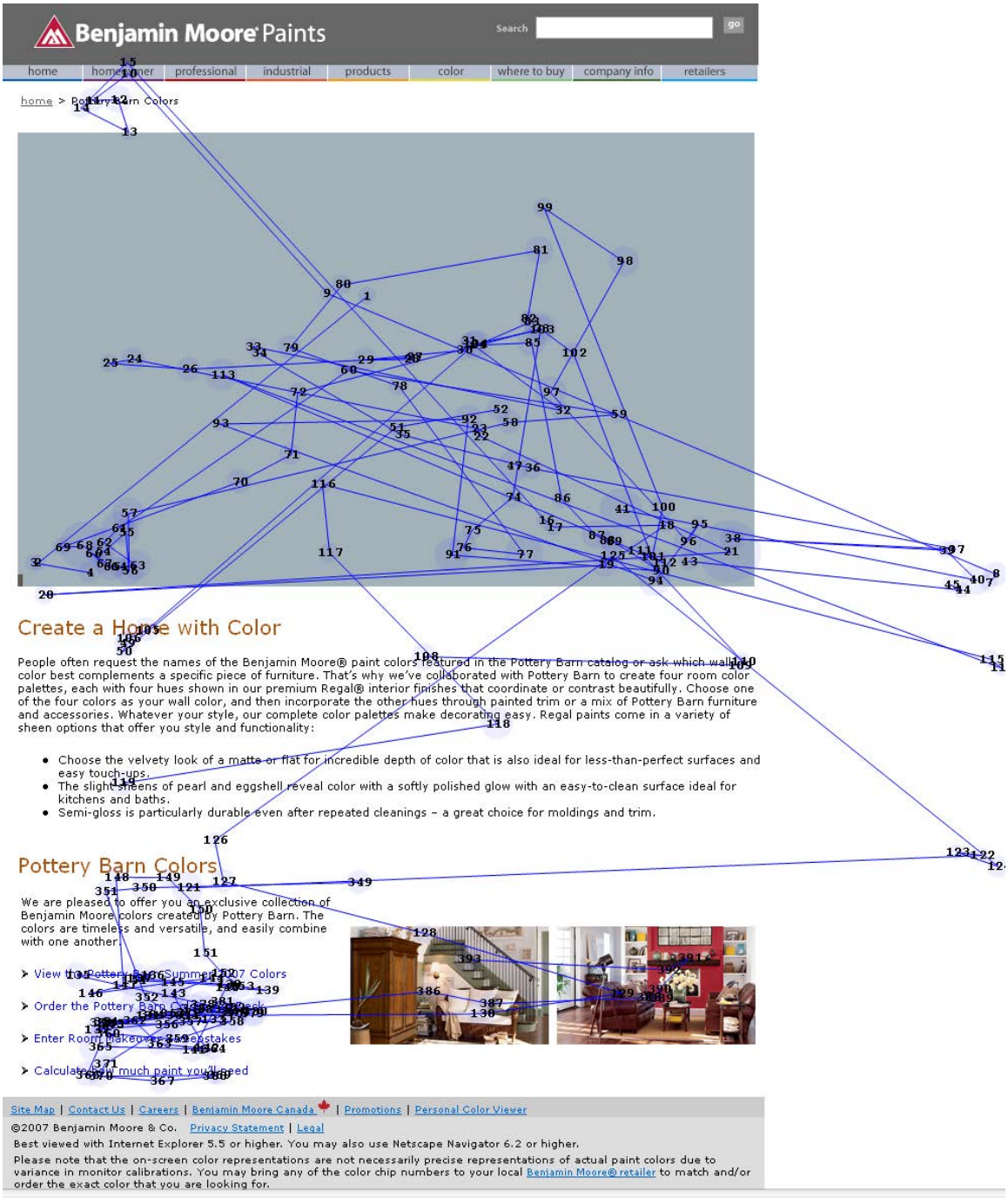
Sometimes part of a heatmap or gazeplot looks off, but the rest seems right. In the following example it looks like the user was fixating a lot on the image on the page and the white space below it. But really he was not.



An example of a misrepresentation on a gazeplot.

In one task users looked at whatever they were interested in on the Benjamin Moore website. Several people were quite interested in paint colors and how they looked in a real room decorated with furniture. Animations and PDF files made the gazeplot of some pages inaccurate. For example, the gazeplot for a page about colors shows a

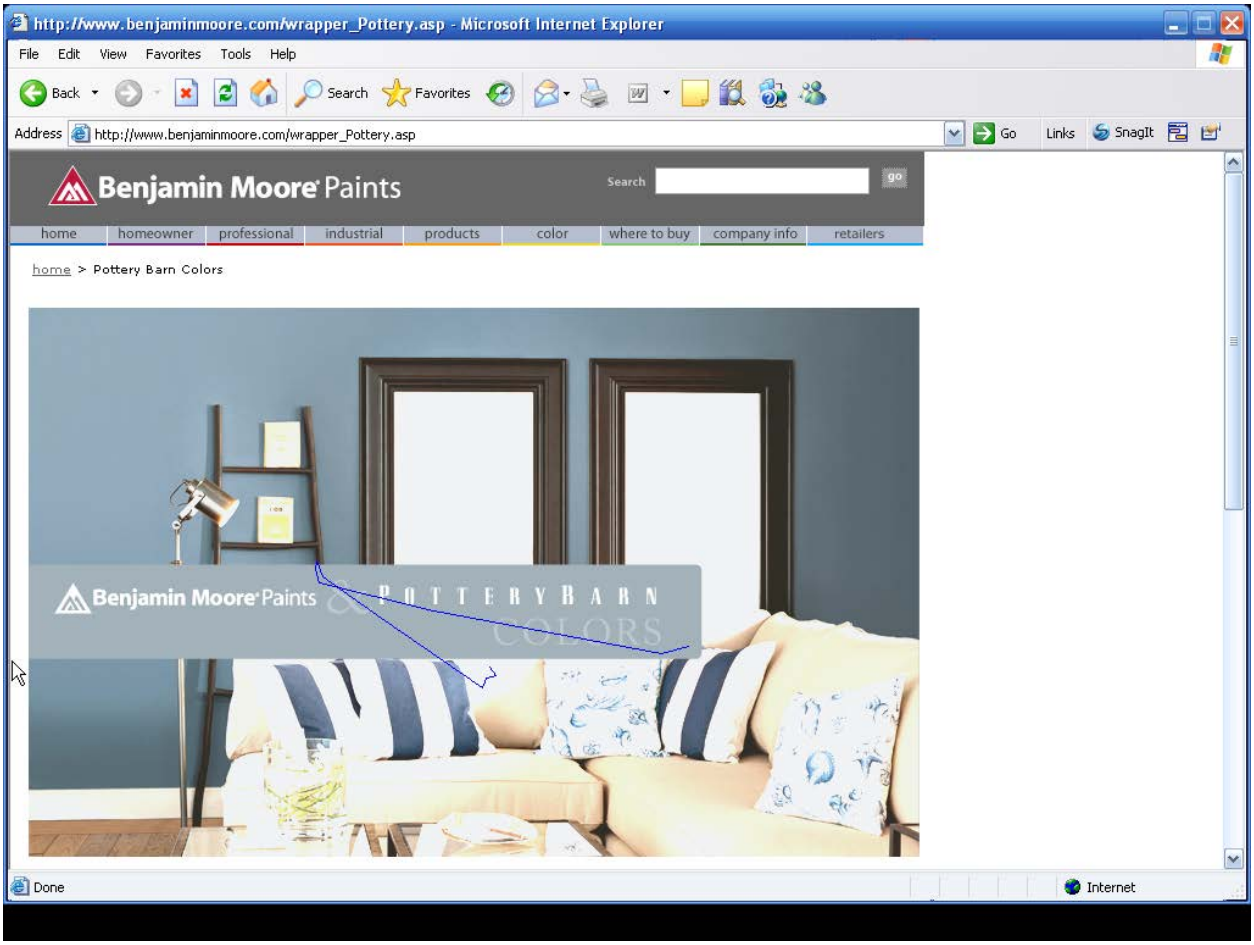
user looking a lot at the plain, grey box. It's too bad we do not know what the user was looking at there.



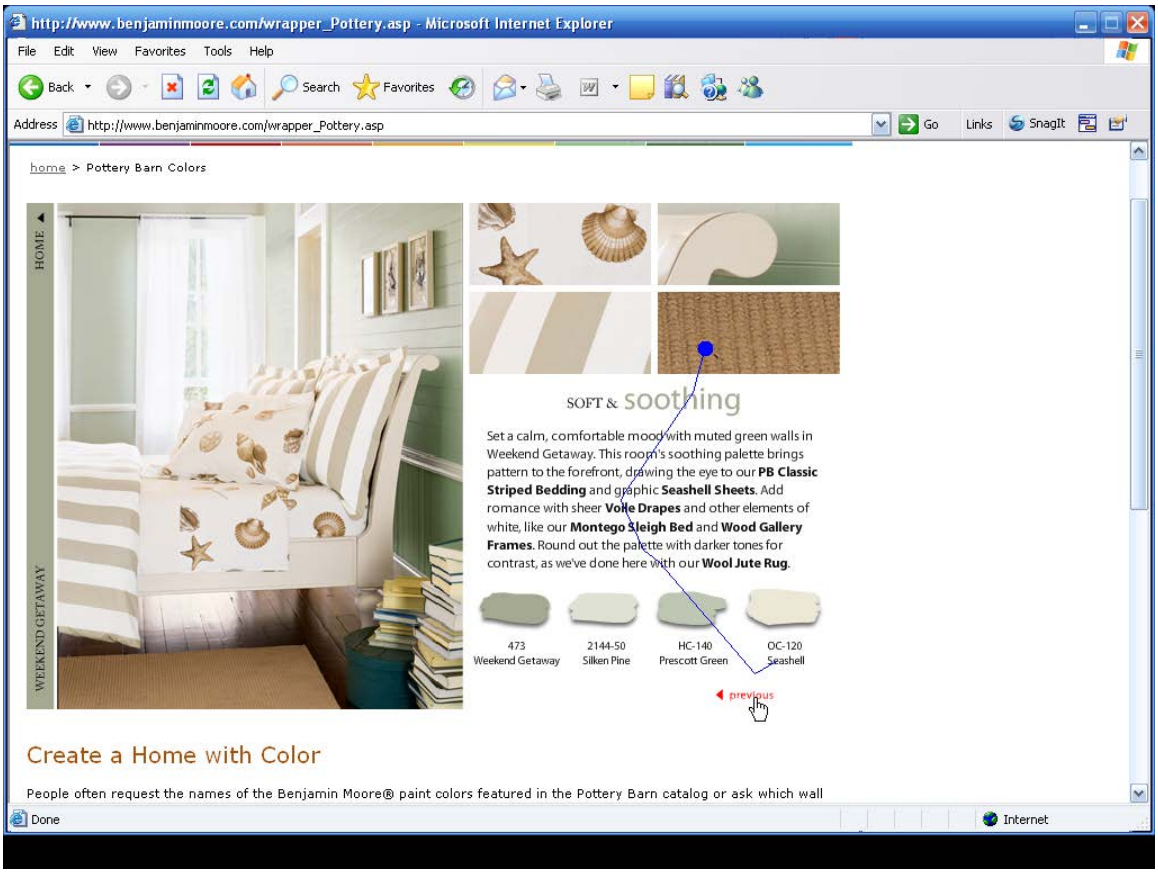
An example of a misrepresentation on a gazeplot on www.benjaminmoore.com.

When looking at the above gazeplot though, you can guess that there is some kind of animation happening, (there was), and go to the gaze replays to analyze it. So this misrepresentation is not the most dangerous. Those that are the sneakiest, depict heat erroneously but in expected areas. So you may never predict that the heat representation is not accurate.

Taking screenshots from watching the gaze replay for this same user and task, we see there were a few animations, as we expected.

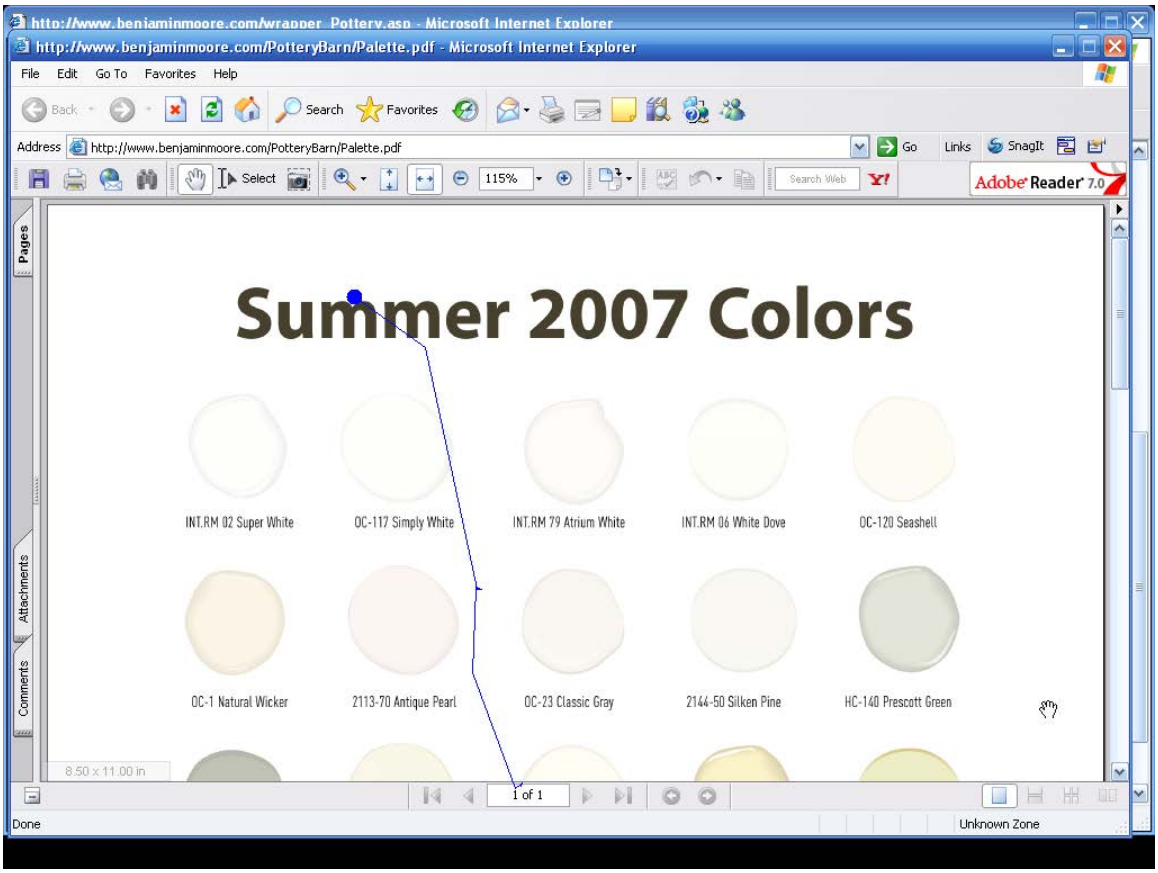


An example of what the user was actually looking at on a page on www.benjaminmoore.com.



Another example of what the user was actually looking at on a page on www.benjaminmoore.com.

The first gazeplot with the empty gray box is a sure indication that there's an animation. But what of the PDF file presenting color swatches? The only way you'd know that this screen was there would be to watch the gaze replays. It doesn't even show up in the list of captured pages, but it was definitely present during the session, as can be seen in the gaze replays. The first image, with the gray box at the top, gives no indication that a PDF file showing paint color swatches was overlaying the page the user was on.



A third example of what the user was actually looking at on a page on www.benjaminmoore.com.

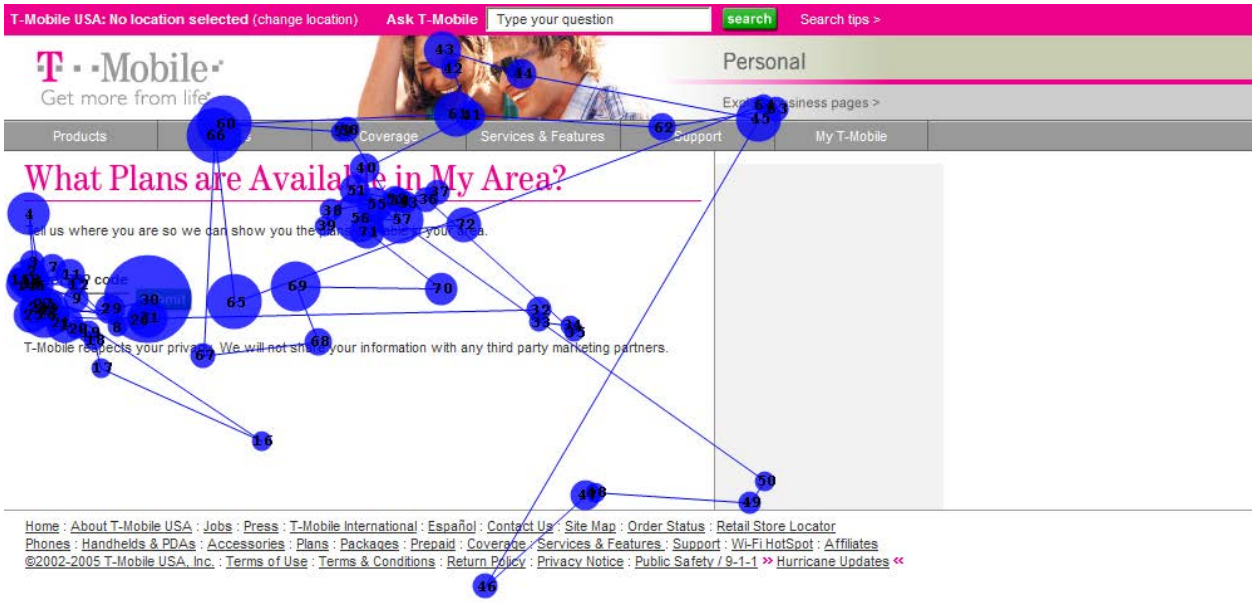
In a similar example, we see a user looking at the *GQ* magazine homepage. It's pretty obvious which things he is interested in. It all seems logical. There are a few fixations in the middle black area that seem right, like looking at the date in the upper right and at the model's picture on the top, but that are really inaccurate.

Most fixations are correct, but some are not. This is a very dangerous misrepresentation because it is hard to detect.

When a pop-up window appears over the main page, it is not accounted for in the gazeplot. But coincidentally the looks on the pop-up were strategically located so they made sense also in the image below it. You'd never know about this misrepresentation unless you watched the gaze replays.

An open window actually got some of the fixations that appear in the gazeplot for the Webpage on www.men.style.com.

This is a dangerous and misleading misrepresentation. In the following examples, there was another window in front that users were actually looking at. It's not easy to tell, without watching the gaze replay, which elements the users were actually looking at.



Another potentially dangerous misrepresentation—this on www.tmobile.com.

These are known issues and ET technology organizations have been working to fix them. Whichever technology you use, we urge you to ensure that these have been addressed in newer versions, or that you account for them in your analysis.

Consider User Recall

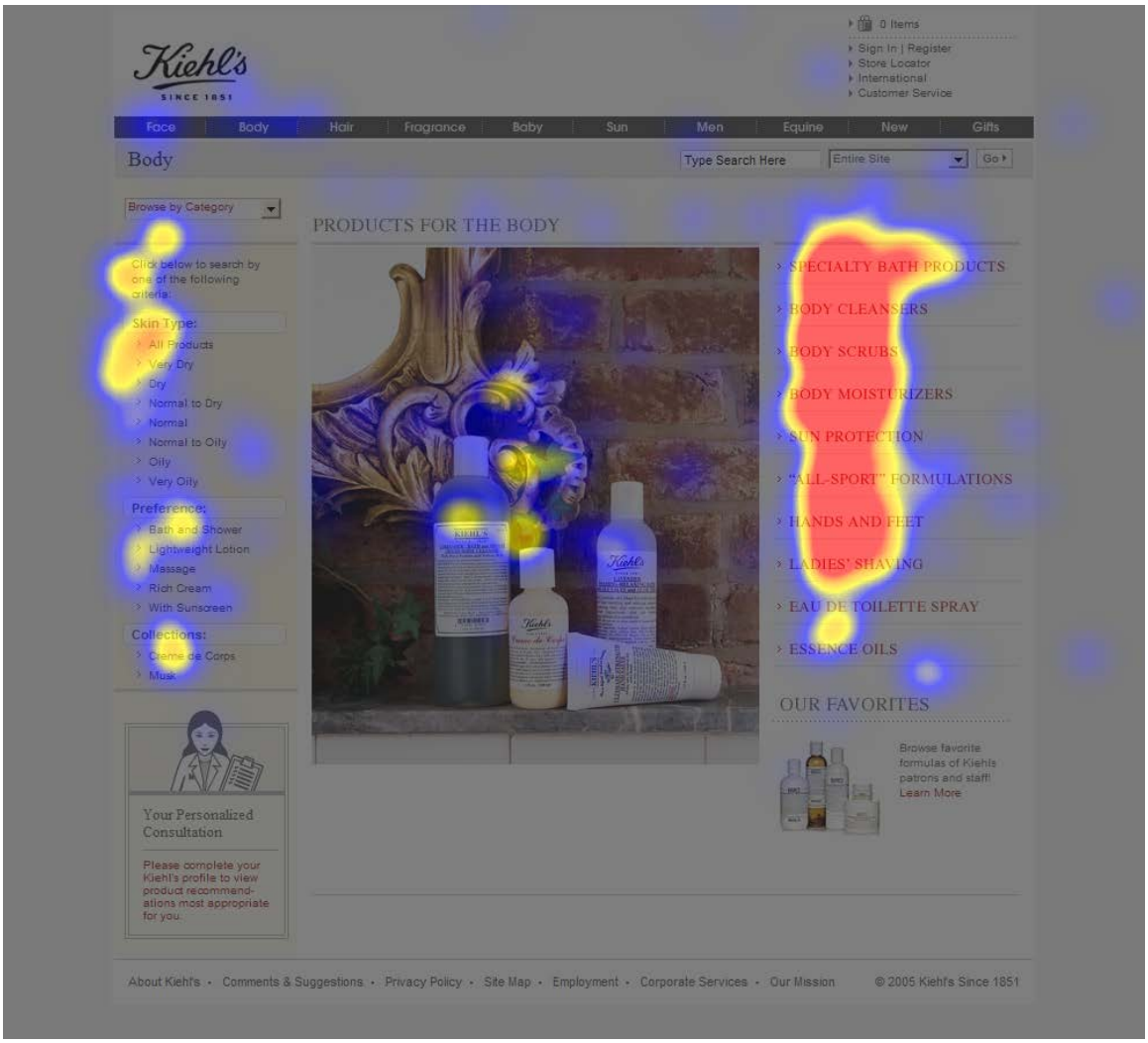
Using the Web is like living life. You don't wake up every day and start from scratch, like a baby (unless you live in the Bill Murray movie *Groundhog Day* or the Drew Barrymore/Adam Sandler movie, *50 First Dates.* There are a lot of movies on this theme, eh?). When a person uses the Web, she is bringing to each page she hits the knowledge that she gleaned from previous experiences, namely, the page she was just on. She remembers what she has already looked at and maybe ruled out. She usually doesn't need to click them again, unless they are incredibly poorly named or the placement of the same items changes, a dirty trick. We refer to this memory the users bring to each page they visit as *user recall.*

***User recall* encompasses the memories that users bring to each page they visit. It impacts where people look.**

One would not expect a user to treat elements that appear on all pages of a website the same the first time she sees them as when she sees those elements the second, third, or fourth time. Knowing this, why would one evaluate one Webpage alone without first considering the whole story, explicitly the pages the users have already explored?

Let's consider the scenario from our research where users were attempting to buy something they wanted on the Kiehl's website. (The company sells body lotions, shampoos, and various other great grooming products.) More than 50 people used this site in our study. When we look at the heatmap from the pages, if reviewed in a vacuum, we could draw some very different conclusions than if we take the whole story into account. For example, take the image below. The users rarely looked at the top horizontal menu. Is it because the choices are not explicit? Or is the menu too small or too grey and people didn't see it? It is because the right-side menu is larger and more attractive than those on the top and left sides?

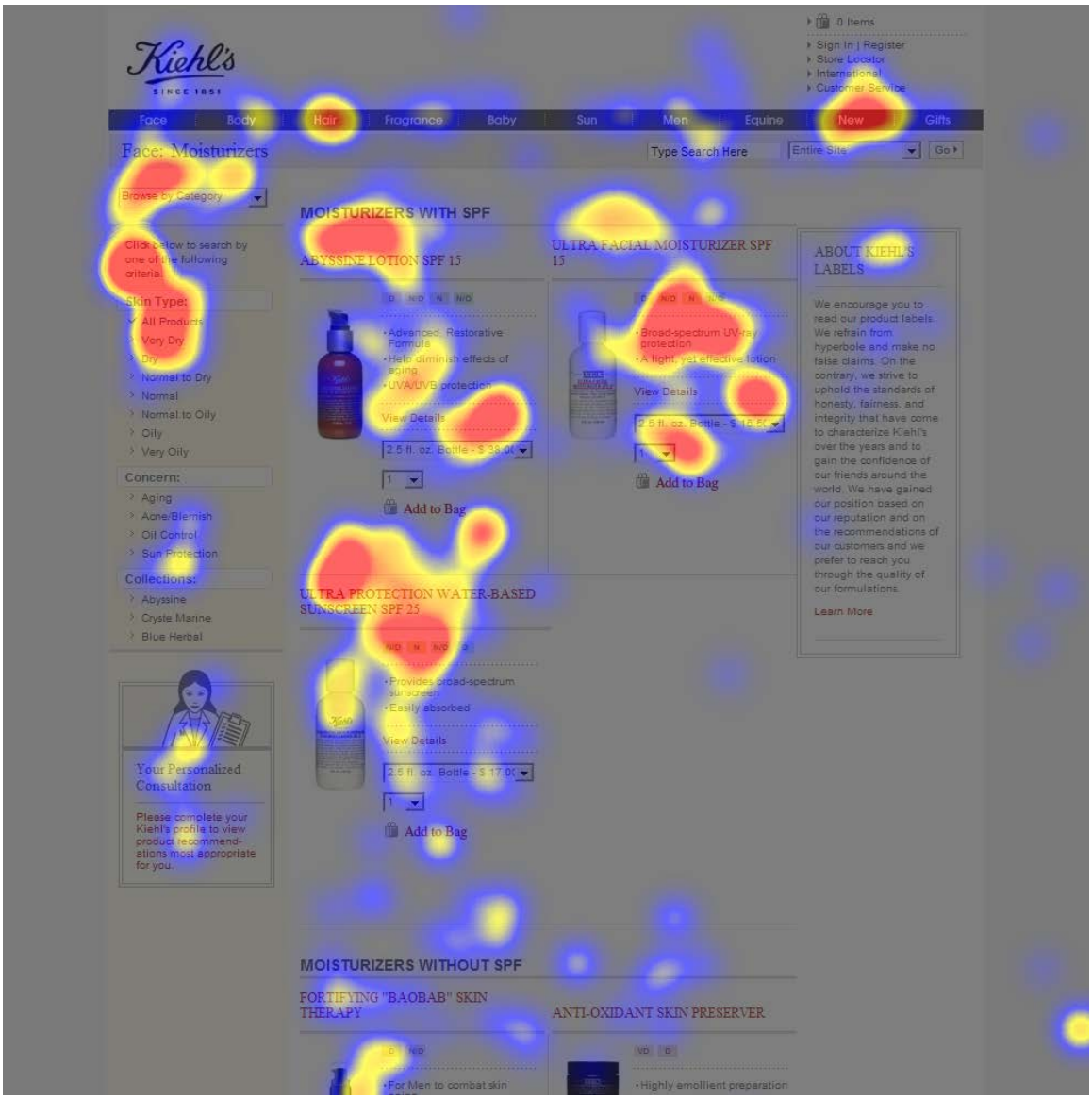
Looking at the heatmap alone one might hypothesize some of these things.



The heatmap shows a lot of activity in the right-hand menu on www.kiehls.com.

But that's not what's going on in this screen. Rather, what happened was that all users had already been on the homepage on the site, and most users had already visited several pages on the site before they ever hit the page above. So prior to hitting that page, we saw users looking more thoroughly at the left-side menu, and a bit more at the top horizontal menu.

While visiting this page, the users had already become very familiar with the site's global navigation. They knew where it was without looking at it, and they had a reasonable idea of the options it would offer them if they were to reach for it. In this case, *not* looking at a screen element is an indication of a good design that follows the usability guideline to encourage consistency. Users practiced *selective disregard.*



This heatmap may indicate that people didn't look at the top menu much. But they did, just not on this page, on www.kiehls.com.

And on the homepage, users looked most at the top horizontal menu. In fact, this is where the user would chiefly decide the type of product he wanted to buy himself. Users really scrutinized these choices for *Face, Body, Hair,* etc. Once this was done thoroughly, they only revisited these top headings when they exhausted looking through the left and right categories, and the products within those categories for their selected genre. For example, one user decided to buy something for his hair, he then looked to the left menu, and to the center of the page, and to the right menu. There was no need to look at the top menu, and he knew it, unless he exhausted looking at all hair products and didn't like any and instead wanted to try, say, products for the face.

Users scrutinized the menu on the homepage on www.kiehls.com.

Heatmaps Do Not Account for Varying Speeds Users Worked At

The different speeds users work at is not taken into account in heatmaps. The heatmaps show the heat for the items looked at no matter when and in which order they were looked at. So if one user was on a page for 30 seconds and another user was on the page for five minutes, they would all still be represented in the same heatmap. If you want to use a heatmap to display the entire visit to a page, but exclude part of the time such as an interview with users at the end of their visit, for example, you cannot do this unless the users took the exact same amount of time to do their task. (An alternative is to interview people in a separate stimulus or file, which we always recommend.)

There is No Depiction of Order in Which Users Looked at Items in Heatmaps

If one user looked at the center of the page first and most of the other users looked at the left menu first, there is no way to get this from the. Even if you did chop off part of the time at the end of the heatmap to remove any strong outlier users, you still would not get any sense of what people looked at first or last.

57. **When using heatmaps, refer to both your facilitation notes and the gaze replays from the sessions. The gaze replays give the best sense of time and order and show even dynamic elements.**

58. **Use individual user gazeplots to display order of fixations.**

59. **Use individual user gazeplots to best display durations of fixations.**

We highly recommend that if you are going to use heatmaps that you refer to both your facilitation notes and the gaze replays from the sessions. There you will see the real apple of the user's eye, like the pop-up that appeared, and not just a snowstorm of white on the homepage.

Fixations vs. Time

It is possible to use heatmaps to visualize different kinds of data from an eyetracking study. Most prominently, one can color-code based on the amount of *time* users spent looking at things or the number of *fixations* users employed looking at things.

In practice, there is rarely much difference between plotting fixation count or gaze duration. In either case, the red areas are the ones that are looked at the most and there are only minor differences in the coloring.

We have chosen to plot all our heatmaps based on the amount of time spent looking, because that's what most people find easiest to interpret.

THE *AREAS OF INTEREST* OR *LOOKZONE* FEATURE

There is another analysis tool in eyetracking technology that deserves a mention here, called *Areas of Interest*, AOIs, or *LookZones*. This feature is meant to help with doing quantitative analysis. The premise is you can draw boxes around elements on pages, name them, then have the system compare and calculate how often and how long these items were looked at. For example, we wanted to know how much people look at images. So, interns went to every single page that every single user looked at and drew these AOI boxes around every single image that appeared. (We also had many subcategories of images classifying the types of people in the images, etc.) Then we exported the AOIs to a spreadsheet to see the total number of users who looked at the "image" AOI and for how long.

AOI / LookZone Limitations

One of the major limitations of this feature is that it does not automatically consider whether the AOI were actually presented to the user. So it will, for example, profess that the user did not look at the image and set the interaction to zero even though the user was not actually ever presented that image and did not have a chance to look at it. All these zeros skew the average.

You can get around this by making AOIs on the whole page and writing a program that considers whether the user was presented with the page and thus the item at all and then calculate the average fixations only including those users who had the opportunity to look at the item. But you have to do some programming work for this.

The other main drawback for using AOIs is the dynamic elements not being considered or captured, which is still a wicked issue.

A final drawback is that it takes an inordinate amount of time to draw all those little boxes. Creating a template can help some with this. But, again, technology is always improving so consult with the technology providers to see if these issues are addressed.

60. When using any quantitative ET features, always test with a very small sample before you embark on a larger study to first ensure that the technology does what you expect it to do.

ANALYZING VIDEO AND ANIMATION

As for getting a static data visualization for video you tested—just forget it. Unless you have a large, highly technical staff who can write a good program to extrapolate this information, expect to have to watch the gaze replays of videos to examine the eyetracking usability data. If you do have a slew of staff or great graduate students to make this program, please let us know when you are selling your program. We'll gladly take a copy.

You can, of course make heatmaps of video, but they are almost always inaccurate, as they only show the static picture of one part of the video, and not all the moving pictures the users looked at. Even in the best case scenario where most of the video is an object mainly standing still, i.e., the "talking head" video, you still get inaccurate reads in heatmaps. And even if you get past this, you can only get accurate information in the heatmap for one user at a time, negating the best reasons for using a heatmap in the first place.

In one example, we made the heatmap below from just one user. Making a heatmap with one user is not recommended for analysis, but we do this here for heatmap discussion purposes. In this image we see the user, expectedly, looking a lot at the face of the man being interviewed. If you were to take this heatmap as is and analyze it, you might come to believe that in this video, the man's face is the most interesting thing to the user, then the tree, then the street sign, then the trashcan. You can easily explain that the user also got a bit bored with the talking head at a few points during the video segment and looked at the links to the right of the video and at the video controls at the bottom. All of these looks are understandable and expected. But what of the looks at the green sign over the man's right shoulder, and those on the tree and trashcan. People do often look at things like signs, so that is possibly a real look and not a misrepresentation. But is a tree or a trashcan really that interesting?

The heatmap from www.cnn.com might lead one to believe that the tree and trashcan were very interesting.

When you actually watch the video you see there are some scene changes, and in fact a large segment shows the same man in a box on the right (the very location where the heatmap made it look like the user was looking at the trashcan) and the studio host in a box on the left. Where the birch tree had seemed so interesting we see it actually was not—the user was instead looking at the small box encasing the image of the female commentator. The image below only shows the time for the image depicted. In other words, we cut out the time with just the man being interviewed, so this heatmap is about right.

The heatmap from another view of the page on www.cnn.com shows that the faces were of interest, not the tree and trashcan.

And when we look at just the man, we see that his face gets the most fixations, then the controls and links, but that the street sign actually is interesting to this user.

This heatmap from www.cnn.com shows only the time when the view was of one reporter. There are more fixations on the face and the sign, and few or none on the tree and trashcan.

Advancements in Analysis of Dynamic Elements

According to Dr. Peter Brawn, the greatest challenge for eyetracking in the field also deals with analyzing the data. Peter explains that the basic output of a mobile eyetracker is a video recording of the scene, with a cross-hair overlaid to display the participant's point of gaze. "The scene is changing all the time as the person moves around and looks at different things," he says. "Compare this to a remote or desktop eyetracker being used for a usability study where the scene (the website) is always appearing in the same place (on-screen.)"

For field studies Peter says that AOI analysis poses even more challenges, "If we take the example of an in-store shopper study focusing on a particular category, then our scene would be an aisle containing a number of products and variants. However, each participant's scene recording is different — a unique video recording as they navigate the store and browse the shelves of an aisle," he says.

So there are basically two approaches for analyzing data for eyetracking field studies, according to Peter, "Use observational coding, which involves manually reviewing the footage and marking time-stamped events as they occur. For example, fixations can be calculated by recording when someone starts and stops looking at something. However, with the eyes moving so quickly and fixations occurring a number of times each second, this often requires slowing the footage for frame by

frame analysis or making multiple parses through. So depending on the complexity of the coding scheme, this could require an analysis effort of 5, 10 or even 15 times the actual length of the footage."

Other analysis solutions allow for dynamic lookzones (i.e., lookzones that can be moved as the objects within a scene move). "However," Peter explains, "multiple lookzones within a moving scene means this also becomes very laborious. For example, as the participant walks along an aisle looking at the shelves all of those lookzones have to move, change size and this has to be done manually. Also, because each participant's scene video is individual, it has been impossible to perform multiple subject analysis in this way."

The good news is that recent advancements have changed the way this analysis is performed and vastly reduced the effort involved. Peter describes a recent advancement in this area, ASL's *GazeMap* software. "GazeMap uses edge detection and object recognition to map environments. For example, you can map aisles within a supermarket and overlay lookzones. The software is able to recognize them from each participant's individual recording and the lookzones adjust in size and position as the participant moves, so that they stay overlaid on the relevant objects. As well as automating this process it allows for multiple subject analysis."

USING THE EYETRACKING ANALYSIS TOOLS IN THE BEST WAY

First of all, while you are doing the testing, always take notes. Don't rely on the gaze replays as sometimes the system crashes and you lose the eyetracking data completely. Also, having your notes can help you sort through the mounds of data you collect.

61. Always take good notes in case the ET crashes and the user's file does not save. Good notes will also help steer you toward particular areas to focus on in the gaze replays.

Once ready to review the data, we found the heatmaps to be a very helpful and useful tool to begin with. We created many heatmaps including all users. Looking through these pages helped us to get some ideas about which designs attracted users and where the problem areas might be. But we did not take the heatmaps as valid representations until we watched the gaze replays or consulted our notes from the sessions. It really is misleading to just blast out a bunch of heatmaps and analyze them alone. There were far too many times that the heatmaps included misrepresentations because of another window popping up or when a user decided to discuss something while he happened to be looking at a page. A heatmap doesn't give you any inkling that this happened. It is just red.

62. Use heatmaps to give you ideas for places to further investigate, aka, what to look for in gaze replays.

After looking at heatmaps, we then employed the gazeplots and gaze replays. In doing this we could conclude which heatmaps were valid and why or why not. And using these tools also enabled us to really see the way items were looked at. This part of the exploration is interesting because it is more qualitative and rich. We were not looking for an aggregate for a specific page, per se. Instead we were looking for patterns. This is necessary if you are going to use the fluid testing method with eyetracking that we used, where we let users move around a website and websites and the entire Web as they pleased. We wouldn't stop a user from clicking a link on

an SERP because 20 other users didn't click that same link. We let them go where they wanted.

If we had prevented users from navigating naturally, we would have recorded unnatural behaviors that had nothing to do with the way real users actually use the Web when they are on their own without a facilitator to restrain them. And what's the use of analyzing data about a behavior that doesn't occur in real life?

With our approach, we could see attributes of pages, even if they were not the exact pages, that emitted the same or a similar response from users. For example, an image with low contrast is relatively easy to qualify, as is an image that is too small for the space allocated to it. If one user was presented with a certain small, low-contrast image and she did not look at it, and another user was presented with a completely different small, low-contrast image and also did not look at it, then we are learning that even in different tasks and on differently-designed pages, people are not looking at low-contrast, small images. We can add these gazeplot cases to our qualitative pile of examples. Technically, of course, the page the image is on affects whether the user will look at the images, as do a number of other factors. But if you gather enough examples of the same reactions to items with the same attributes, the confidence level in the hypotheses rises significantly, until you feel sure enough that it is theory. And you can use many gazeplots in this exact way.

Tasks Discussion

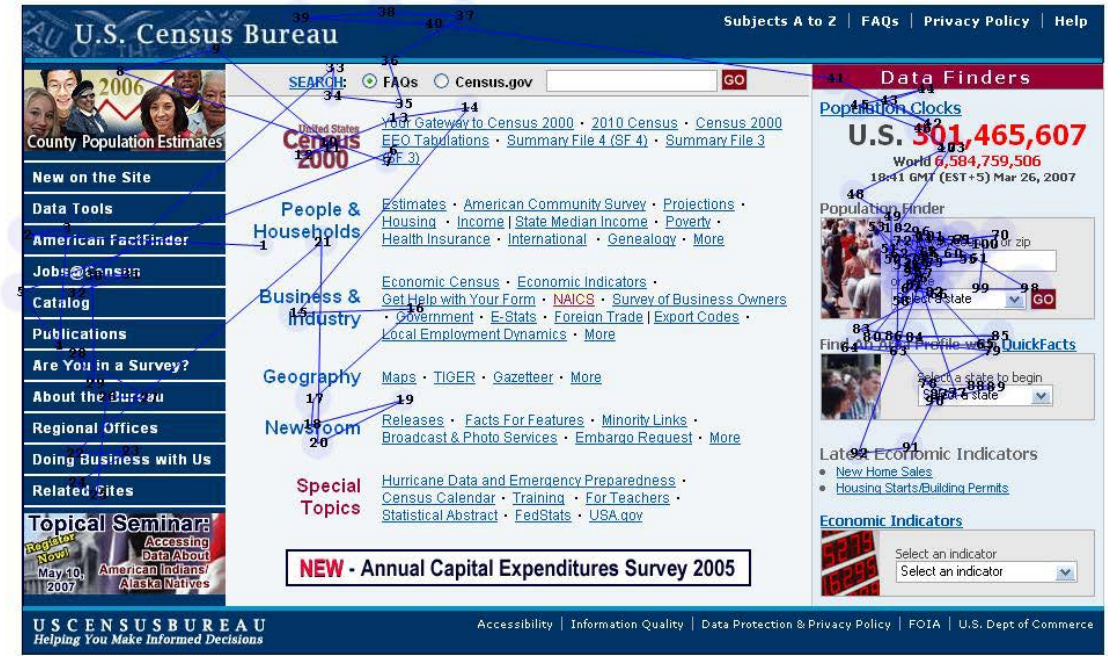
THE *SCENIC VIEW:* SHOW ONE PAGE AND INTERVIEW METHOD

Unfortunately, because of the way eyetracking technology works, it is a lot easier to run a test on a few Webpages than on many Webpages. And it is even easier to show the user just one page. Additionally, the mountains of data are smaller and easier to analyze when you test one Webpage at a time. So in some studies, the research leader will set up the eyetracker so a particular Webpage will appear for the user. The user is not able to move around the website, clicking links and seeing pages as a reaction to the links he chooses to click. Rather, the user is slowly led through pages at the test facilitator's guidance. This is really not at all like reality on the Web. The Web is a fluid medium. A user's reaction to a page is highly related to the link he clicked to get to that page. Links set expectations.

Navigation and movement are the lifeblood of the Web user experience. It is crucial to allow freedom of movement in a usability study for the results to have any validity and for your conclusions to help you make your site more profitable when real customers go shopping. Solely testing individual static pages without navigation is exactly the same approach as that taken by the proverbial drunk who looks for the car keys under the streetlamp despite having dropped them on the other side of the street. You may have an easier job looking, but you won't find anything you can use.

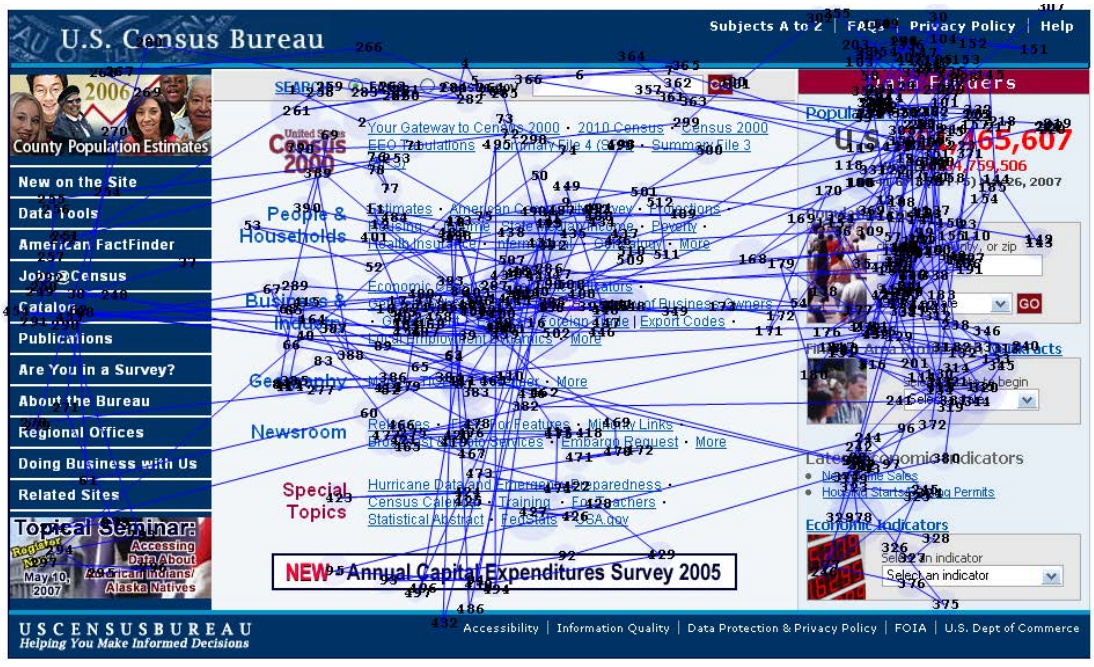
Also, people do not typically gaze all over a page while they are really trying to get work done. They look for key words and phrases then bam—off to the next page. But when a user knows he is only getting one page and he cannot click and work, he will look more at the one page. (Note, you may tell a user he has some task to do and he will try to click on the first task, but he will quickly learn that the test process is: 1) facilitator opens a page, 2) I can comment on it but cannot click at all. So the user's mindset is more on studying and commenting on the page rather than rapid scanning and clicking to get the job done.

In one example, we see a user looking at the United States Census Bureau homepage. When trying to do the task of finding the U.S. population and that of the state of Texas, the user looked at the menus, the main content headings, and mostly at the population clock area on the right.



While trying to find population numbers (i.e., doing a realistic task) on www.census.gov, the user looked at the tool on the right, a few headings, and the left-side menu.

But when asked to discuss the same page, the user looked at much more and for much longer.



When discussing the page on www.census.gov, the same user looked at a lot more. (But this behavior would not happen during real-world use.)

63. Asking a user about a page rather than having him do tasks is not recommended usability methodology and is poor eyetracking

usability methodology. You must be diligent with your usability methodology when giving tasks for the user to do on one page.

But consider typical, non-eyetracking usability testing with prototypes. Often you only do have one page designed and want some feedback on it. You can certainly collect good data about this page. But we firmly believe in taking a task-oriented approach to research even if there is just one page. Asking the user to do a task is good behavioral research because it simulates reality and enables the designers to see how the person would actually work with the design. Asking the user what he thinks about a page you put in front of him is more suited for something like advertising research done in a focus group.

64. If testing just one page, communicate the task (if there is one) to the user, then launch the page—in that order.

If testing just one page:

- Do ask the user to do a task, then launch the page (or ask him to go to the site if you must actually bring users to a particular site).
- Don't open the page, let the user examine it, then give him a task. Doing these steps in this order would encourage a user to examine the page when he might have never done this on his own. This is problematic in non-eyetracking usability testing and is particularly problematic in eyetracking studies.
- Don't open the page and ask the user to "Please look at the page and tell us what you think." This truly encourages the user to explore a page in a way they probably never would. Most people are trying to get something done when they go to a Webpage. That something may be finding out the latest information on the website—it may be that open-ended. But even that is directed in some way.

In managing the usability program for a software company, Kara as often as possible had a line-up of designs that the development teams needed feedback on. The less revenue-generating products got lower priority and sometimes were awaiting help. So when a user came in for a specific lab test, if there were even as few as ten minutes at the end of the session, she would get feedback on one of the items paused in the queue, provided the participant matched the target user profile for that particular design/product. From the user's perspective, if he expected to be in the lab for two hours and we were finished with the first product's testing in an hour and forty-five minutes, it's fine to add on another task or two if the user does not seem to be fatigued, and you are careful to end the session at the time you originally agreed to. This way, you as a facilitator can chip away at the lower-priority items that need some testing. You can also save some money on the honorarium payments you make to users by distributing these micro-tests across tests that are already scheduled for some other product. The main drawback is some developers who work on these items cannot make it to the lab on short notice. So, try to give them a warning if a slot for testing their item may be opening. Another possible issue is that if users do not typically finish the main test early then these items in the line-up may not get any feedback for a very long time. A final shortcoming is that even if the user is not actually fatigued, he may have some trouble switching gears to the new

product/design if he has been strongly focused on the same product interface for close to two hours.

TASK-BASED TESTING

Why Give Tasks in Typical Usability Tests?

Whether to give tasks in usability testing has long been discussed in the usability field. Some contextual inquiry purists believe that users should be allowed to do anything they choose. One problem with this is that in a lab setting sometimes users don't want to do anything. We have tried this in the lab and some users really just have no ideas about what they could do. They sit there and say, "I don't know." This is pretty awkward, and you end up giving them tasks anyway, just after embarrassing them a bit.

It is certainly valuable to ask users when you are scheduling them whether there are things they might need to do on the Web during the time of the study. Then you can watch them do their personal tasks. Or sometimes when we do tests at a user's office or home, even if we do give tasks, at the end of the session we will ask the user if there is anything he needs or wants to do. Sometimes he will go back to one of the websites we tested and try something there. Or he might do a task he had been wanting to do earlier in the day. This can be very revealing. But it is not completely realistic in a development environment to expect that users will choose, on their own, to test the areas that you are currently designing, interested in, and need usability feedback on. It is very difficult in practice to hit all the areas of the user interfaces that you want to hit if you don't direct the user in some way to them.

Also, if the user has no goal in mind, be it his own or one we provided for him, there will be no way to judge whether the person had success in using the website. Anyone can just look at Webpages and click around, saying, "That's nice." But what about when they get home and actually try to get something done? Not watching people attempt real tasks can leave big holes in user research when doing product design.

The main reason to base usability tests on tasks is that this best mirrors the way people actually use the Web: there's a reason you visit a website. People have to make an active determination to go to any specific website, and thus they have some kind of goal in mind. Sometimes this goal is very specific (say, buy flowers and have them shipped to Mom for mother's day). Other times the goal can be more vague (say, check if there's a good movie playing tonight). Sometimes, the user's goal is even to kill time (say, play a computer game or read the main news headlines). In all cases, however, there is something the user wants to accomplish and that caused the user to visit this particular site out of the hundreds of millions available on the Internet.

This "pull" users experience is quite different from the way other media work. For example, a magazine is a "push" media form that arrives in your mailbox on a schedule determined by the publisher: you pick it up when you see it. (Then, of course, you may put it down and only read it when you feel like getting the kind of information contained in that magazine.)

In any case, more than a decade's worth of experience shows that users are highly goal-oriented when they visit websites, so that's the scenario we have to establish for usability studies if we want them to be realistic.

Supplementing Task-Based Testing

There are so many reasons to do task-based testing in a controlled lab environment. But thorough development teams will supplement this research with good field studies and other usability research methods such as remote testing, card sorting, prototype testing, and heuristic reviews.

Now we consider giving users tasks in eyetracking sessions, specifically. The reasons for giving users tasks are the same in ET sessions are they are in non-ET sessions.

The effect the task has on the eye gaze, when a user is doing a task is an interesting concept. Naysayers argue that giving the user a task had an effect on what he looked at on the Web. To this we say, yes, of course it did. When you are trying to get something done, you look for words, links, commands, images, and information that will help you get to your final destination. And if that destination is different from the last, you will look for and take a different route. What you are trying to do surely affects what you look at.

But giving a user a task is warranted, the main reason being: People do tasks on the Web. In fact, that's all they do. Stuff. Even in our contextual field studies we can confidently say that users hardly ever go to a Web page just to look around and contemplate the design. They always have some task in mind, even if it is very open-ended, like finding out what a site is about, or reading the news. For example, even on a news site a user may give himself the task to look for something interesting, or what's happening in politics, or gossip, or who won the game last night. These are all tasks. These are things users do. People don't hit a corporate website page just to look around. Typical users just don't have that kind of time. There is laundry to do, sales to make, people to see.

Designers do look at websites though. We have to. We all survey other websites to see what kinds of design elements they are presenting and how. Even when we go to sites to do a specific task, we find ourselves checking out other areas just to see what the designers did. This is not normal. This is a behavior of Web designers, not Web users.

To ensure that giving a task did not have an adverse effect on the users' eye gaze, we did several very open-ended tasks and allowed them to do what they wanted. And in some cases we told the users we would open several websites, saying something like: "When the first opens, if you are interested, use it. If not, just let me know and we will move on to the next." In many cases users took a look and said, "not interested" and we were true to our word and moved on. But in these scenarios where the user was interested in the website we opened for them, users did not just look all over the page and inspect every crevice. Instead they created their own tasks almost instantly, looking at the *part* of the site (e.g., a breaking news story) that particularly interested them.

VARIETY IN TASKS

When giving users tasks in a usability study it's beneficial to include a variety of types of tasks. None of them should be overly leading. Of course, a lab study is known for being far more leading than a method such as a field study where you observe users doing whatever they want using their own tools in their own space. It would be especially interesting to do an eyetracking study at the users' sites and allow them to do whatever they want. Transporting and setting up the eyetracker would be pretty inefficient however. Even in a lab study and while giving the user tasks, the types of tasks can range from very to far less directed. In our study we gave a wide variety of tasks. (Our tasks are discussed in the "Our Research" chapter

of our book, *Eyetracking Web Usability*, by Jakob Nielsen and Kara Pernice, due late 2009.)

Tasks That Require an Answer but no Particular Website Need Be Used

In order to have some end point to the tasks, and to determine whether a user was successful, we asked users to do tasks that were mostly open-ended but that called for an answer. In order to record and start the eyetracking technology using the settings we needed, we opened the browser for the user. But you could start them on any page, even a blank one.

One example of a task is, "What is Skype and is it something you might want to use or not?" We started users on a blank page for many tasks. And in some cases we started them on some news site such as *The New York Times'* site. We opted for this on occasion over a blank page because we thought it might be interesting to see if anything on the news site was interesting enough to distract the user away from his task. Sometimes users would see some new item that caught their eye, but more often than not they went straight to the URL field and typed Google.[7]

The reason we did not start users on the Skype site, in this example, was because we wanted people to do the research how they chose to. For this same reason, we did not open Google for them or have the Google toolbar available in the browser.

Tasks That Require an Answer and Start on a Particular Website

In some cases we wanted the users to work on a particular site. And if you don't open the site for the user, or at least tell her to use the site, she will often end up on some other site to find her answer. The benefit to having the participant actually use the site we want to look at outweighs the main drawback: Bringing the user to the site would boost his confidence level in using the site. For example, if you ask a user to find out how to tie a bowline knot, and you bring her to the Boys Scouts of America website, even if a knot page is not immediately visible the user may try harder to find it than she would if she had come across the website herself. Deliberately bringing the user to a site can lead her to believe that the information she needs must be on the site, somewhere. But if we just tell the user to find the answer and bring her to a blank page, she searches on Google (or however she wants to) and clicks a hit, having no idea whether the site she clicked on has the information she wants. She may then be more likely to click *Back* and check out a different site that Google suggested.

It hardly needs stating that this behavior is the one that occurs in real life and therefore the one we want to replicate in our research for the findings to be as applicable to real sites that want to maximize their value to the business. It does no good to design for the scenario where the user knows that your site has the answer to a problem, with a very small number of exceptions. (For example, an online banking site might be designed under the assumption that existing customers would

[7] Obviously, people could use any search engine they wanted, and some users went to another engine than Google. It is not our job to do market research to establish the relative market shares of different search engines, but in our study, the vast majority of users went to Google when given a choice. In the future, another search engine may become the predominant choice, but this is not likely to change any of the findings about what people do once they arrive at the destination site. All that's important to know is that new visitors often arrive from some search engine where they were looking for something specific.

expect the site to be able to state their account balance, because people know in advance which bank has their money.)

Rapid-fire Tasks

We wanted to get information about where users look first for some standard UI elements on a Webpage, such as the main menu, the search box, the shopping cart, the privacy policy, and the jobs area. We thought it would be interesting to mention the area name, such as “the search box,” then launch a website for the user. This way the user knows the task he is trying to complete before the Webpage launches, so every fixation is likely focused on the goal of finding the aforementioned item.

Any Interest? Tasks

Arguably, bringing users to a site and asking them anything at all is giving a task. But we kept some of our tasks extremely open-ended. In fact, the task typically was something to the effect of: I am going to open a website. If you are interested, go ahead and use it however you want to. If not, just let me know you are not interested. Many users were not interested and would say so. When they were interested, they did all kinds of different things.

Should You Run Your Own Eyetracking Research Study?

For the vast majority of companies in the world, the answer is easy: No, you should not conduct eyetracking studies yourself.

The reason for this advice is that most companies have never even done a simple user test of their website, intranet, e-mail newsletter, or software applications. For any of these design problems, you can get a huge amount of usability improvements from the simplest of studies that you can run in a few days. No reason to go to the extra expense of eyetracking.

The concept of "low-hanging fruit" is a very tired cliché, but it's nevertheless true when it comes to usability. Any design that has never been subjected to a systematic usability evaluation will be so riddled with major design blunders that the smallest usability project will suffice to produce a long list of urgent redesign recommendations.

In fact, if you have never done any usability evaluation, you can start with something even simpler than testing: just check your site for the guidelines we have pointed out in our eyetracking book and report. (Also read our other books and reports for even more guidelines—hint, hint :-))

In general, though, we do recommend testing because no published guidelines can address all the individual issues that are specific to your industry and your business. Usability guidelines can only tell you about those design issues that are common for most websites (or intranets, or e-mail newsletters), and so they provide a starting point for improving your site, but they don't have the final word. Only your own customers do.

So yes, you should test your site with real users. But do you need to know exactly where their eyes point at any given time during the study? No. All of the big usability catastrophes that cost you millions of dollars in lost business are obvious to an observer who sits next to the user, watches where they click, and listens to what they say as they think out loud.

A user might say something like "I expect to see shipping costs before proceeding through checkout. There's no way I will type in my credit card number without knowing the total I will be charged." It's pretty obvious that you need better disclosure of the shipping costs to improve your conversion rate. (In fact, an e-commerce site might easily double its sales from this simple fix to the user experience.)

In most cases, the shipping costs are probably nowhere to be found on the page, so it's obvious that you need to show them. In a few cases, the shipping costs might actually be there, but the user simply didn't see them. Eyetracking might help you in this latter case, but you still wouldn't know exactly how to redesign the page layout. All you really know is that the old design doesn't work and is costing you lots of lost sales.

The recommended approach is to mock up one or more alternative designs as paper prototypes and test them with a handful of new users. How do you get the budget to bring in five more users? By not having spent your money on an eyetracker.

65. Don't be too quick to start using eyetracking. As a precursor, skip the ET and instead mock up one or more alternative designs as paper prototypes and test them with a handful of new users.

Even companies that have done user testing usually haven't done enough of it to exhaust the rich mother lode of insights that can be mined by simple, cheap studies. Just get hold of five representative customers, make them do a few representative tasks, and ask them to think out loud while you stay quiet. Keep doing this again and again, for all the different parts of your site and all the different kinds of things you want your customers to do. Keep redesigning the site to fix many usability problems you will invariably find. And then test again, because even an "improved" design will be so far from perfect that there will still be loads of further improvements to be found through the simplest of tests.

Only those companies that have a highly evolved usability culture and that have run hundreds of simple tests should even remotely entertain the possibility of doing an eyetracking study. Yes, you will find new things. And yes, as our research has hopefully shown, the eyetracking analysis will give you new insights into those issues you already knew about. But from a cost-benefit perspective, you don't need this added level of refinement until you have already covered all the basics with much cheaper methods.

ROI is the name of the game for usability. The goal is to maximize the profitability of your website. Eyetracking has much lower ROI than other usability methods, so it should be pretty far down on your to-do list.

WHY NIELSEN NORMAN GROUP INVESTS IN EYETRACKING

Considering that we are recommending that most companies do not spend any or considerable resources on eyetracking and instead spend on lower-cost iterative methods, why did we conduct several large-scale eyetracking studies ourselves? (In addition to the studies documented in this report, we have also written a book and various reports based on eyetracking usability research on reading, searching, email newsletters, and e-commerce sites. All can be purchased on our website, www.nngroup.com.) All told, we have spent more than half a million dollars on eyetracking studies and the analysis of eyetracking data.

First of all, we wanted to learn about how people look at Web pages in more detail than we have done before. And, we wanted to test some of the 2,397 usability guidelines we have produced over the years and see how eyetracking refutes or supports them.

Second, we did say that eyetracking is an option for companies with a highly evolved usability culture, once a company has run hundreds of regular non-eyetracking studies. This description certainly fits our own company: we specialize in usability; we invented many of the most commonly used usability methods; and we have conducted endless usability research in the past (at the time of this writing, Nielsen Norman Group had tested 1,217 websites systematically, to which should be added the several thousand sites we've tested informally during open tasks where each user could go where they wanted, and many more our team members worked on before joining NN/g). Because we already know so much about Web usability, it was worth the high cost for us to delve even deeper so that we could generate even more insights for our usability conference tutorials and consulting projects.

Third, we have to admit that one reason we embraced eyetracking was in order to write several reports as well as our newest book. Eyetracking provides a new way to illustrate Web usability that's very suited for publishing a full-color book that's going to look good in bookstores. We hope you agree that this report as well as our book and other research reports have provided an interesting, fresh look at Web usability findings.

Fourth, a major part of our business is to produce an annual usability conference series, *Usability Week*. Gaze replay videos from the eyetracking studies are a great way to illustrate many of our usability points in a way that's more engaging for the conference audience than simply showing screenshots. If people are going to sit through a week's worth of usability training seminars, you need to keep them motivated.

These last two reasons don't apply to most companies: You're probably not about to write your fourth book about Web usability, so you don't have to seek out new material. And most companies are not in the business of teaching other companies how to do usability. Fixing the company's own site is more than enough work.

Still, a bit of our reasoning may apply to your situation: Eyetracking is showbiz. It's a way to get colorful pictures to sell a book and engaging video clips to sell a seminar. By the same token, eyetracking provides much needed razzmatazz to impresses clueless people on your team or in your management who don't understand usability. Spending your company's money to bamboozle the bosses can be a good investment, if that gets them to follow your usability recommendations, even though you could have derived the same insights at 20% of the cost without the pretty pictures.

Technology-Related Notes

Eyetracking technology has come such a long way in terms of testing PC software and websites. Gone are the days of users having to put in special contact lenses or wear headsets. These were quite invasive in what could have already been a nerve-wracking experience for some people. Can you imagine? A user comes to the lab, and you tell him:

> I will be watching you closely, as will fifteen other people. I will be taking notes, giving you tasks to do. And we'll be watching exactly what you are looking at. Oh, and can you put on this ridiculous headband (or hat). Yes, like that. Oh, no. No. You look good. It becomes you, really.

We recall taking a tour of a lab at a large organization that had employed eyetracking usability for several years. Researchers had the head-camera device. To "make it fun" they had the ET camera affixed to a baseball hat. Our colleague's reasoning was the baseball cap was less invasive and intimidating than another make-shift device to make the camera hold fast to the user's forehead. Yes, the baseball hat was better than duct tape around the head. But the users were not comfortable with it; the camera shifted around a lot; and forget the (OK, we'll say it) possible vile contagious items you can contract by sharing headwear. No thanks.

The Tobii monitor has changed this and is nothing short of revolutionary in terms of recording users in a relatively unobtrusive, transparent, and yes, hygienic, way.

That said, these are a few of the pitfalls that relate to ET technology that we recommend you beware of:

- We lost about one task per session of 10 tasks. No recording, just a crash and no save. This is expensive and annoying. (They were always the good tasks too!)
- Sometimes the system crashed so hard it took an inordinate amount of time to reboot. Allow the users to get some coffee or take a bathroom break if this happens.
- The system can be slow to load and save files.
- You cannot run more than one process at a time. For example, if you are exporting a file, that's all that the ClearView software can be doing. You can't make the most of your time by reviewing some gazeplots or even creating new stimuli.
- The system is surprisingly bad at giving feedback. For example, when exporting a file it never gives any sort of progress indication or even that it is working. You just wait at its mercy until either the "complete" or an "error" message appears.
- The above is particularly frustrating since exporting larger studies takes a long time, measured in hours.

- We did get a lot of error export files and still don't know why. We'd just reboot everything, sometimes unplug, and it would miraculously work. We were constantly saying things like, "Don't touch it! Don't make it angry."
- Files take up a lot of space. We were constantly moving files off the PCs and onto thousand-dollar external hard drives. (Certainly these get cheaper every day. But you still have to plan to pay for *huge* data storage.)

As noted, Tobii has many fabulous traits, one being the people who work for this Sweden-based company. They have made great strides and continue to work hard to improve the system.

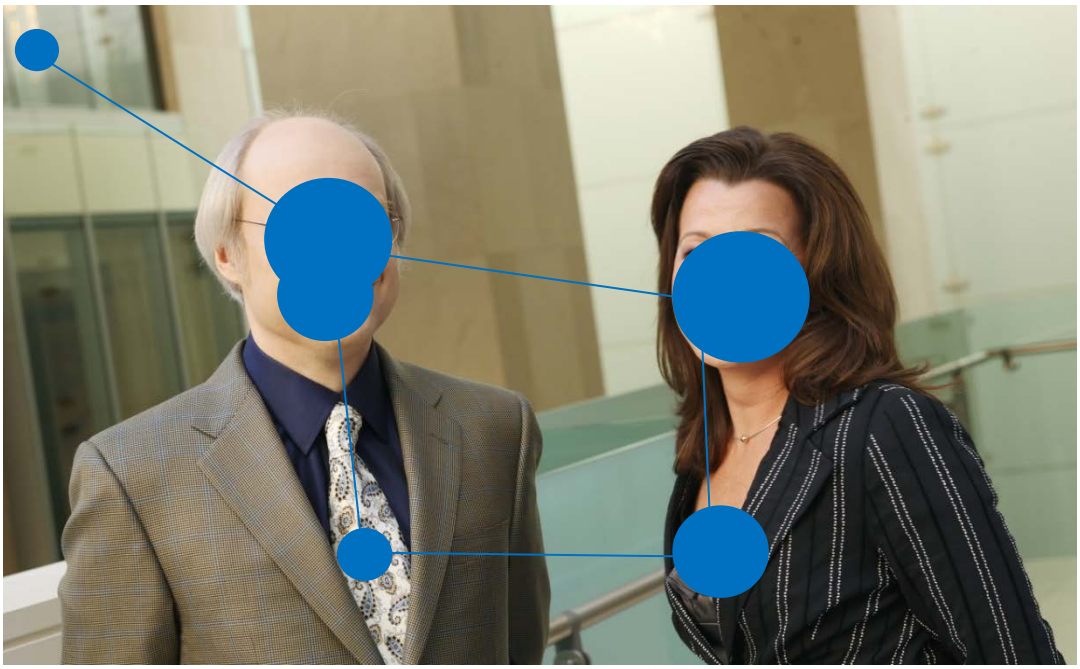
About the Authors

Kara Pernice is the managing director at Nielsen Norman Group and heads the company's East Coast operations. Since joining NN/g, Pernice has led several intercontinental research studies and written associated reports about topics such as intranets, the Web and accessibility, senior citizens and the Web, public relations, and site maps. She has developed and taught numerous seminars about these topics and about a variety of usability methods such as the product life cycle, field studies, emotion and design, usability testing, and eyetracking. Before joining NN/g, she established successful usability programs at Lotus Development; Iris Associates, an IBM subsidiary; and Interleaf. She managed the first usability program for Lotus Notes and the Domino server, and after her team's work, *PC Magazine* wrote, "If this were summer camp, Lotus Notes would walk away with the Most Improved Camper award." Pernice chaired the Usability Professionals' Association conferences in 2000 and 2001, was presentations chair for UPA 1999, and was conference advisor for UPA 2002. She has an MBA from Northeastern University and a BA from Simmons College.

Jakob Nielsen, Ph.D. is a principal of Nielsen Norman Group. He is the founder of the "discount usability engineering" movement, which emphasizes fast and efficient methods for improving the quality of user interfaces. Nielsen, noted as "the world's leading expert on Web usability" by *U.S. News and World Report* and "the next best thing to a true time machine" by *USA Today*, is the author of the best-selling book *Designing Web Usability: The Practice of Simplicity* (2000), which has sold more than a quarter of a million copies in 22 languages. His other books include *Hypertext and Hypermedia* (1990), *Usability Engineering* (1993), *Usability Inspection Methods* (1994), *Multimedia and Hypertext: The Internet and Beyond* (1995), *International User Interfaces* (1996), *Homepage Usability: 50 Websites Deconstructed* (2001), and *Prioritizing Web Usability* (2006). Nielsen's Alertbox column on Web usability has been published on the Internet since 1995 and has about 200,000 readers. From 1994 to 1998, Nielsen was a Sun Microsystems Distinguished Engineer. His previous affiliations include Bell Communications Research, the Technical University of Denmark, and the IBM User Interface Institute. Dr. Nielsen holds 79 United States patents, mainly on ways of making the Internet easier to use.

The book, *Eyetracking Web Usability*, by Jakob and Kara, was published late 2009.

Jakob Nielsen and Kara Pernice



Acknowledgements

Thank you to our extremely hard-working research assistants who toiled to make our studies possible:

Sukeshini Grandhi for many hours testing in the lab, recruiting, and doing pretty much anything else we needed.

Tatyana Nikolayeva for creating websites, coding analysis tools, and drawing AOIs.

David Shulman for great efforts on analyzing the data and drawing countless AOIs.

Many thanks to Susan Pernice for so much work in recruiting, finding locations, drawing AOIs, and assisting with the data analysis spreadsheets.

Also thank you to the good people at Tobii Technology for answering so many of our very odd questions and doing so with a smile.

Thank you to Rob Stevens, Peter Brawn, and Andrew Duchowski for sharing their eyetracking knowledge with us.

Finally, thanks to Jacqueline Holland for copyediting this report.

NN/g

Evidence-Based User Experience Research, Training, and Consulting

Since 1998 Nielsen Norman Group has been a leading voice in the user experience field.

- Conducting groundbreaking research
- Evaluating interfaces of all shapes and sizes
- Guiding critical design decisions to improve the bottom line

We practice what we preach

We don't just talk about the importance of testing with real users, on real tasks, in real life business situations: we do it. Every week, somewhere around the globe, NN/g team members are conducting research that informs the three pillars of our business: training, consulting and research. In that work we have:

- Tested over 2,000 different interfaces
- Observed more than 4,000 users—in person—in 18 countries and on 5 continents
- Analyzed thousands of hours of recorded user observations sessions
- Conducted countless diary studies, focus groups and remote user tests

Our collective experience will save you time… and money

Making technology easier to use is no longer a nice-to-have. Useful, usable products make money. And our expertise can help your team achieve their design goals quicker and easier than going it alone. Choosing NN/g means you benefit directly from our:

- Finely tuned methodology: We have an arsenal of proven tools at our disposal and know how and where to apply each one, taking the guesswork out of how to achieve the optimal design solution to meet your business goals.
- Comprehensive body of knowledge: We've taken the results of our decades of research and testing and distilled it down into actionable guidelines, best practices and proven methodologies. Our research library, containing over 50 published reports, X books and an email newsletter archive dating back to 1995 is unrivaled.
- Practical approach: Our approach is 100% practical, useful and actionable. Whether you attend one of our Usability Week events or invite us to consult at your place of business, the training you will receive can be put into action immediately so that you can see the results.

Our people are the best in the business

At NN/g there is no "B Team". When we dispatch consultants to work with you and your team, or when you attend a Usability Week course, you are learning directly from some of the best-educated and most experienced minds in the business.

- Our principals are considered pioneers in the fields of user research and interface design.
- Our researchers and consultants tackle the most recent and relevant topics in usability, from evergreen challenges such as information architecture and intranet usability to emerging trends in social media and mobile usability.

Stay Informed

Jakob Nielsen's Alertbox Newsletter

Summaries of our latest research and insights published twice per month.

To subscribe: http://www.nngroup.com/articles/subscribe

NN/g

Evidence-Based User Experience Research, Training, and Consulting

TRAINING

Usability Week Events

Usability Week training events are offered in the U.S., Canada, the U.K., Europe, Asia and Australia.

Each week-long event features full-day, immersive training courses where attendees learn practical skills directly from experienced practitioners so they can solve complex UI problems and create better interface designs.

Over 40 courses offered in these categories:

- Agile
- Applications
- Content Strategy
- Credibility & Persuasion
- Email
- Information Architecture
- Interaction Design
- Intranets
- Mobile & Tablet
- Non-Profit Websites
- Prototyping
- Social UX
- User Testing
- Visual Design
- Web Usability
- Writing for the Web

Available courses and upcoming locations: www.nngroup.com/training

In-house Training

Many of our courses can be taught at your location and customized to fit your unique offerings, methods and resources.

In-house training is ideal for:

- Large teams that want to spread user experience perspective throughout the group
- Teams working on large projects that need to kick start the creative process and head in the right direction

In-house training information: www.nngroup.com/consulting

NN/g

Evidence-Based User Experience Research, Training, and Consulting

REPORTS

NN/g has published over 60 reports that detail thousands of evidence-based design guidelines derived from our independent research studies of websites, intranets, application, and mobile interfaces.

Over 60 reports addressing these topics:

- Agile
- Applications
- Audience Types (e.g., children, college students, seniors, people with disabilities)
- B2B Websites
- Corporate Websites
- Ecommerce
- Email
- Information Architecture
- Intranets
- Mobile & Tablet
- Non-Profit Websites
- User Testing
- Social UX
- Strategy
- Web Usability

Shop for reports here: www.nngroup.com/reports

CONSULTING

The same experts who conduct our research and teach Usability Week training courses are available for custom consulting including:

- **Evaluating your website, application, intranet or mobile interface** (average cost $38,000 USD)
- **Usability testing** (average cost $35,000 USD)
- **Strategic planning** (average cost $12,000 USD)
- **On-site training with your team** (average cost $9,000 USD per day)

Consulting details: www.nngroup.com/consulting