Stephanie Y. Hsieh. Comparing First Impressions of Display Clutter Between Cultures. A Master's Paper for the M.S. in I.S degree. May, 2017. 54 pages. Advisor: Robert Capra

Clutter is an aspect of web aesthetics that has come to the fore in recent years as the research community works toward a fuller understanding of what clutter is and how it affects users' perceptions and performance with interfaces. In this paper, I examine the possible effects of cultural background on users' first impressions of display clutter in website designs. The study was conducted using a series of five-second impression tests that asked participants from two distinct cultural groups to view a set of screenshots encompassing multiple levels of clutter, as measured by a JPEG file size measure. Results showed some effects from cultural background on perceptions of clutter, raised some issues with the cross-cultural applicability of the objective JPEG measure's ability to predict subjective judgments, and provide some evidence that organization is a key distinguishing factor between visual complexity and clutter.

Headings:

Display Clutter

Cross-Cultural

Usability

Web Aesthetics

First Impression

# COMPARING FIRST IMPRESSIONS OF DISPLAY CLUTTER BETWEEN CULTURES

by
Stephanie Y. Hsieh

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

May 2017

Approved by

_____

Robert Capra

## Table of Contents

## List of Tables



## List of Figures

## Introduction

The increasing integration of the Internet into daily life around the world has made research into web design an ongoing concern. In particular is the nebulous concept of display clutter. This idea, described by Moacdieh and Sarter (2014) as “a widely acknowledged but ill-defined problem” (p.61) is an important area of research for several reasons. From a human factors and ergonomics perspective, there is the fact that display clutter negatively impacts user performance (Donderi and McFadden, 2005; Harper et al. 2009; Moacdieh and Sarter, 2014). From a design perspective, “clutter” is a term often applied to interfaces that users find unappealing and therefore don’t wish to use or interact with in the first place, regardless of actual usability (Lindgaard et al., 2011; Tractinsky et al, 2000). Donderi and McFadden (2005) explored both of these relationships, measuring levels of clutter using the compressed file size of JPEG screenshots. As Donderi and McFadden (2005) note, this method of using the JPEG file size is an established way of quantitatively assessing relative levels of clutter in an interface. In their study, Donderi and McFadden (2005) demonstrate that the JPEG measure is related to human performance navigating visual displays, as well as subjective, aesthetic judgments of those same displays. This connection between quantitative measures and subjective reactions, and how aesthetics affect the latter, is one of the major issues at the core of this study.

At the heart of this connection is the fact that, while “clutter” has quantitatively demonstrable effects on individual's reactions to and ability to use visual interfaces, the

same term has also entered colloquial usage. However, as the general public becomes increasingly accustomed to using visual interfaces such as websites, and becomes more sensitive to design and usability issues that affect their daily life, it becomes increasingly important to understand how to connect subjective, but truthful, criticisms with well-defined concepts and quantitative measures so that the objective and subjective effects of display clutter can be better understood.

The quest to understand the concept of display clutter is a difficult and complicated one. Among other issues, it's quite possible to train people to tolerate levels of information density and/or display clutter that would otherwise be extremely difficult to process. Furthermore, there is the possible role that culturally determined web aesthetics plays in determining what kind of web design standards a given individual is used to or expects—regardless of how well that aesthetic relates to a design's actual usability. As such, any understanding of measures of display clutter must not only be able to account for the physiological limitations of humans' ability to process information, but also the way that subjective reactions can vary due to culture and training.

All of this coalesces in the particular subject of this study, namely: to understand how subjective responses to varying levels of display clutter vary between groups of different backgrounds. Our particular area of interest is in how an individual's cultural background affects their reactions to display clutter. The importance of investigating the cross-cultural dimension is supported by recent research that has turned up evidence that cultural background may in fact affect a person's sense of web aesthetics (Cheng and Nielsen, 2016; Cyr et al., 2010; Reinecke and Gajos, 2014). Such studies recognize that,

in an increasingly global world, understanding how to accommodate different cultures' expectations in interface design is important for producing globally usable products.

## Related Work

Central to this study is the concept of clutter in visual displays such as websites. While 'clutter' is a term that is frequently used in casual conversation, attempts to pin down and define its meaning for research purposes have encountered many difficulties. As Rosenholtz et al. (2007) states: "we lack a clear understanding of what clutter is; what features, attributes, and factors are relevant; why it presents a problem; and how to identify it" (p. 1). Nearly a decade later, this problem persists in the research community, as Moacdieh and Sarter (2014) cite Rosenhotlz et al. (2007) in their 40-page review of attempts to define and measure display clutter, concluding that "the concept [of clutter] is too broad and ill defined" (p. 94). This review of related work will start by touching on that very problem, going over some work in the area of defining clutter that is particularly relevant to this project. Then we will discuss the relationship between clutter and usability, followed by a discussion of the relationship between clutter and web page and display aesthetics. Finally, this review will discuss some recent research into the cross-cultural dimension of usability in web interfaces.

### Defining Clutter

#### Cluttered vs. Complex

Attempts to define what cluttered means in the domain of visual interfaces very quickly run into the problem of differentiating what we call 'clutter' from the related

concepts of visual complexity. All of these are strongly related concepts that revolve around the amount of 'stuff' present on a given interface—images, text, headers, etc. How do we understand these concepts in relation with each other, if they are indeed separate concepts at all?

Working with photographs in an effort to determine what perceptual factors viewers use to determine the level of complexity present in an image, Oliva et al. (2004) finds that "perceived complexity of an image…depends on the amount of perceptual grouping" (p. 1042) which implies that the *organization* of the 'stuff' on an interface may be the key to differentiating between what is cluttered and merely complex. This is borne out by Oliva et al. (2004)'s experiment, where two groups were asked to sort photographs based on level of visual complexity. Both groups received the same instructions:

> "Visual simplicity is related to how easy it will be to remember the image after seeing it for a short time. Visual complexity is related to how difficult it will be to give a verbal description of the image and how difficult it will be to remember the scene after seeing it for a short time." (p. 1043)

However, the experimental Structure group received some additional instructions. They were told that: "Visual complexity is related to the structure of the scene and therefore, is not merely related to color or brightness" (p. 1043). Between the two groups, a much higher percentage of people in the Structure group mentioned 'clutter' as one of the criteria they used to sort the photographs by visual complexity. This finding suggests that the level of organization or structure in an interface's design impacts whether a visually complex image is considered complex or cluttered. Oliva et al. (2004) conclude that "visual complexity is principally represented by the perceptual dimensions of quantity of objects, clutter, openness, symmetry, organization, and variety of colors" (p. 1044). By this, they posit that cluttered interfaces could be considered a subset of visually

complex interfaces, making any study of the former necessarily a study of the latter. As this study is focused on clutter and not on visual complexity as a wider construct, the terms will be used throughout the review in the following way:

- *Clutter* refers to 'cluttered' interfaces
- *Visually complex* refers to interface that are considered complex but *not* considered to be cluttered

These definitions do run into two further problems as 1) these definitions are idiosyncratic to this review by necessity, due to a lack of consensus definition; and 2) this review has not yet discussed *how* to define clutter in and of itself, beyond Oliva et al.'s suggestion that organization is a distinguishing factor. There is little to be done for the first point than to stick to the nomenclature described above and to note throughout the review where researchers' definitions diverge or coincide with those definitions. The second leads us handily to the next topic of discussion.

### Measurement and Evaluation

Having considered the place of clutter in contrast to related concepts, how do we go about defining or measuring levels of clutter in a given image (including images of an interface, like a static wireframe or screenshot)? This is a tricky position to be in, since 'less organized' is a rather vague way of measuring levels of clutter. Over the years, there have been many attempts to create a way to definitively measure and define visual complexity, clutter, or both. Stickel et al. (2010) suggest an algorithmic way of calculating visual complexity in websites using static screenshots and a simple mathematical formula they call the XAOS metric. Throughout their work, they treat visual complexity as a concept in opposition to visual simplicity and make no attempt to

distinguish between ‘good’ complexity (what would be ‘visually complex’ by our definition) and ‘bad’ complexity (what would be ‘clutter’ by our definition). This makes discussing their work somewhat tricky, though in truth Stickel et al. (2010) don’t use a definition of visual complexity that is in opposition to this review’s term definitions. For the purposes of discussing Stickel et al.’s work, we will use visual complexity throughout, but note where their definition of visual complexity could map onto our definition of clutter. So, to continue: Stickel et al. (2010) use a method of definition visual complexity that is based on: the number of possible interactions, number of visual gestalt groups, and summed entropy of RBG values. This relates to three of the factors Oliva et al. (2004) put forth as perceptual factors of visual complexity, namely: quantity of objects, organization, and variety of colors. Notably, organization is one of the factors that came up as a possible key distinguishing factor between ‘clutter’ and ‘visually complex.’

The emphasis on interactivity is new to Stickel et al. (2010), however. Interactions are calculated by counting up functional elements, or “all kinds of links and active GUI elements like buttons, drop boxes, checkboxes, etc.” (p. 284). This introduces a dimension to the evaluation of visual complexity that is *not* dependent on user gaze or even actual user interaction, but is rather focused on the number of available *potential* interactions as identified by the designer. By introducing this element, Stickel et al. (2010) make visual complexity not just a static quality of an interface, but one that is intimately related with a user’s projected interaction with an interface. Stickel et al. (2010) note that in empirical testing, their XAOS metric matched subjective user ratings of visual complexity when both were applied to the same images, suggesting that there is

some merit to considering potential interactions as part of how complexity is calculated or perceived.

Also of note is the fact that the XAOS metric succeeds in combining several of the measurement perspectives described in Moacdieh and Sarter's (2014) review of extant clutter defining and measurement techniques. The XAOS metric combines the Display Density Perspective, where clutter is viewed as "a surplus of items or objects within a display," and the Display Layout Perspective, described as one where "poor display organization, structure, and order are the main factors that are assumed to result in display clutter" (p. 62). These combined approaches can complement each other, resulting in what is potentially a richer understanding of what factors influence an interface's level of visual complexity.

Rosenholtz et al. (2007) include a similar element of interactivity when defining clutter, though their definition is based, not on the number of potential interactive elements, but on overall user performance. Clutter is defined as "the state in which excess items, *or their representation or organization* [emphasis added], lead to a degradation of performance at some task" (p. 3). This definition is not in opposition to the definition of 'clutter' that we defined earlier, but it does expand the scope of the concept. The Rosenholtz et al. (2007) definition again highlights organization as a key defining factor of clutter, something also noted by Oliva et al. (2004). However, user performance is a new factor and something that adds great complexity to the task of measuring clutter.

The XAOS metric created and put forward by Stickel et al. (2010) was a static measure that could be applied to an interface by simply examining it objectively and plugging numbers into a simple mathematical formula. User performance, when it

entered into the definition, was based on projected assessment of the effort required to use the interface rather than objective measurement of said effort. Rosenholtz et al. (2007) make understanding—and measuring—clutter a much more intensive task by entwining user performance even more deeply with straightforward examination of a visual display. How do we define 'good' performance and measure its degradation? How do we account for factors unrelated to the display that may nonetheless affect user performance? Extant usability testing techniques are ideal for answering these questions, but they also make the work of identifying levels of complexity and/or clutter much more time consuming. For example, someone familiar with e-commerce sites will know that there is always a shopping cart icon where a user can review their selected purchases to date. Someone with similar technological competence but no experience with e-commerce may find the task of checking their selected purchases to be more difficult due to this lack of knowledge. This is just one way in which non-display related factors such as a user's prior experience or domain knowledge could affect user performance with an interface. This perspective goes against what Moacdieh and Sarter (2014) refer to as "view[ing] clutter to be primarily a display property, that is, a function of the features and elements of the display" (p. 64).

However, this is not the only method of measuring clutter (or visual complexity). The XAOS metric is also not the only way of approaching the mathematical or objective means of measuring or defining image complexity. Donderi (2005) notes that "JPEG and zip compressed file sizes predict the subjective complexity of images generated by marine electronic displays, websites, and nature photographs" (p. 823)—an assertion based on earlier work in this domain. The JPEG measure is a simple way of measuring

levels of visual complexity by taking JPEG screenshots of interfaces of interest and comparing file sizes. The fact that the JPEG measure is correlated with subjective perceptions of visual complexity suggests that it is an easy and potentially useful shorthand for levels of complexity in an image.

However, Moacdieh and Sarter (2014) note that image-processing based methods still have a ways to go before they are truly useful. While they are practically quite useful for quickly providing clutter or complexity assessments for many images, a simple score—especially one produced by an algorithm as basic as JPEG compression—can't distinguish between visual *complexity* and visual *clutter*. While the number can provide a measure of how much 'stuff' is on a given interface in relation to JPEGs of similar image quality, this measure can't distinguish between well-organized or poorly organized stuff. Also, Moacdieh and Sarter (2014) caution that context is a significant influence on any correlation between algorithmic clutter assessments and performance data. While, as Donderi (2005) notes, the JPEG measure correlates with subjective perceptions of visual complexity, this finding doesn't necessarily translate to any objective measure of actual performance.

As for Moacdieh and Sarter (2014), the perspective taken throughout their review is one that strongly emphasizes usability: "the presence of performance and attentional costs that result from the interaction between high data density, poor display organization, and abundance of irrelevant information" (p. 65). Their definition again highlights the importance of organization as a key distinguishing factor between what is complex and what is cluttered. It also highlights a certain tension between complexity and clutter, by suggesting that there is a need to ameliorate the effects of clutter by

carefully controlling, not just how elements are arranged on a web page and how many elements are included, but also *which* elements are included. As with Rosenholtz et al. (2007), this leads to an expansion of the definition of clutter, as least when viewed from a display-only lens. An emphasis on user performance, while placing greater demands on the assessment side, is nonetheless useful for grounding all the display- and design-related factors like choice and placement of elements in something concrete and measurable that also (unlike the JPEG measure) allows researchers to distinguish between what is cluttered and what is merely complex.

**Clutter and Usability**

The link between clutter and usability is an interesting one. This relationship has been the site of study as an evaluation method, as the clutter definitions proposed above (Rosenholtz et al, 2007; Moacdieh and Sarter, 2014) demonstrate. Focusing on user performance as a measure of visual complexity is also what Moacdieh and Sarter (2014) refer to as the Performance Evaluation Methods of assessing display clutter.

An exemplar of a study that focuses on this assessment perspective is when Donderi and McFadden (2005) demonstrated that compressed file length (i.e. the JPEG measure discussed above) can predict a user's search time and number of errors when searching a visual display. Search time and number of errors are classic ways of measuring performance in visual search tasks. By tying these performance measures to algorithmically generated clutter measurements in the form of JPEG file size of each screenshot of interest, Donderi and McFadden (2005) aim for an objective, convenient, quantitatively based means of predicting user performance by trying to correlate objective performance with the JPEG measure.

Harper et al. (2009) make the implication in Donderi and McFadden (2005)'s work explicit by presenting a definition of complexity that suggests visual complexity as an implicit measure of cognitive load. Their findings conclude that some of the elements users use to distinguish between simple and complex web pages include elements such as text, menus, links, lists, images, ads, etc. and how those are arranged and organized on the page. Interestingly, their findings also suggest that the amount of information available 'above the fold' on a web page is a key factor—"It is important to note that simple pages fit on to the screen with no scrolling required" (p. 10). Pages designated as visually complex have a "main factor that is always present…the diversity of information, which requires the user to constantly switch context" (p.11). Here the tie to cognitive load appears, as it seems that users approach large, complex pages with an understanding that the larger the page, the more effort will be required to understand the information presented upon it and act.

This is quite similar to the interactive element counting used by Stickel et al. (2010) in their proposed XAOS metric, in that an increase in the number of potential interactions seems to correlate with a perception of complexity in a web page. The key difference, however, is that Stickel et al. (2010) go in and count up all the elements, presumably with an expert's knowledge of what is and isn't actually interactive, whereas Harper et al. (2009)'s finding is based solely on user *perception* of possible interactions. Regardless of the actual number of interactive elements, it is the number of potential interactions that the user perceives at a glance that go into their designation of whether a page is complex or not. Harper et al. (2009) highlight this difference in their findings, noting that:

> "Objects that made pages more complex were not necessarily highly visual components but those that signified the possibility of increase[d] cognitive load… Visual complexity seems to be an implicit key into the perceived cognitive load of the page and the interaction that the users think will be required to use the resource." (p. 14)

This is an interesting contrast to more ergonomics-focused attempts to define clutter, which are dependent on objective measures of performance (Donderi and McFadden, 2005; Rosenholtz et al, 2007; Moacdieh and Sarter, 2014). The implication of Harper et al. (2009)'s findings is that simply looking at a web page in static form—their study was conducted using printed screenshots—leads to users forming conclusions about what interacting with that page might be like, were it a live site. Therefore, the look, feel, and design of a web page may be an important factor in shaping users' opinions of pages, whether they would be easy to use, and whether or not they would like to continuing using them—all completely independent of *actual* performance on a given page.

**Usability and/or Aesthetics**

In Harper et al. (2009)'s study, participants were given free time to examine and rank web pages according to levels of visual complexity. This is certainly important and instructive for understanding how users define the concept of complexity and clutter, but not exactly ecologically valid. People generally do not spend long amounts of time judging the possible effort required to interact with a website; they glance, form an impression, and then either continue their interaction or move on to a site with a more palatable design. Lindgaard et al. (2011) describes a study centered on users' trust of health information websites. Some sites were quickly rejected as information sources based on a first impression, which was described as serving as a "screening device" for users browsing for information. Reasons for rapid rejection included "cluttered, complex

screen layouts, boring use of color, and other design-related issues" (p. 3). Interestingly, this suggests that to the users' eye, there may not be a difference between 'clutter' and 'visually complex,' even if there appears to be a move toward strong differentiation in how research in this area is framed. Overall, Lindgaard et al. (2011) report that aesthetically pleasing design informed users' snap decisions about whether a site was worth trusting and spending time on.

The link between aesthetics and user perceptions of usability has been explored elsewhere and along dimensions other than trustworthiness. Tractinsky et al. (2000) describes aesthetics and usability as "two orthogonal dimensions of HCI" (p. 128): the former is subjective and cannot be quantified, while the latter can be examined using relatively objective and quantitative measures. This seems intuitively true; it isn't hard to think of examples of beautiful design elements that make websites harder to use, such as inappropriate uses of HTML-defined headers. But is this true from a user perspective? Evidence points to 'no,' as Tractinsky et al. (2000) found that user perceptions of interface aesthetics strongly correlated with user perceptions of the entire system's usability. The initial aesthetic-based assessments of a system's usability persisted even after actual interaction with the system—even when the system proved to be less usable than anticipated.

The influence of that initial aesthetic impression is a strong and persistent one. Lindgaard et al. (2006) points out that the power of the first impression has a long-standing history in research literature across many disciplines, citing studies from the realms of marketing, psychology, and neuroscience. In an effort to actually quantify how quickly users form that important first impression, Lindgaard et al. (2006) narrowed the

time spent forming a first impression to 500ms. Despite giving their participants no more than 500ms to view a page and form their first impression, they found that participants formed remarkably consistent perceptions of visual appeal, even when participants were given more time to closely examine the design of the pages used in the study. That initial half-second was enough time for participants to decide which pages they liked and which ones they did not. Similar 'first impression' tests in the form of 'five-second test' have long been used in industry to test the initial impact or appeal of designs. In these applications of the test, a static web page is displayed for five seconds, followed by a few open ended questions designed to gather the participant's impression of the web page. Gronier (2016), studying the validity of this test from a practitioner perspective, found that five seconds was enough to "create a user experience for hedonic aspects. And…the elements that determine the ease of use (usability) are considered" (p. 19).

Overall, research suggests that there is a link between perceptions of aesthetics and perceived usability, and that this perception of usability based on desirable aesthetics is formed within the first few seconds of viewing a web page. This first impression can even persist after actual interaction with a web page, even when said interaction results in a less-than-usable experience. Aesthetics therefore have a role to play when it comes to users' opinions of a system, for good or ill. As Tractinsky et al. (2000) notes, "[t]he façade of an information system is what users experience first and it is what cues users about the inside. Moreover, the façade taints how the user perceives further interactions with the system" (p. 140). The use of the word "taints" is an interesting one in this context, as it implies that this influence may not necessarily be to the benefit of users. The degree to which aesthetic pleasure can overcome usability issues is outside the scope

of this review, but notably, Cheng and Nielsen (2016) found that "Chinese and Western users experience the same difficulties with complex sites, but Chinese…prefer fairly high information density" (Introductory summary paragraph). This suggests that there is a basic level of usability that must be met for all users, regardless of background, and while aesthetic preferences may flavor an individual's perception of usability, said preferences may not affect the objectively determined usability of an interface.

**Aesthetics Between Cultures**

In recent years, the question of research generalizability has come to the forefront of academic debate. As Sturm et al. (2015) point out, very little of the world's population is WEIRD (from a Western, educated, industrialized, rich, and democratic society), yet much HCI research—indeed, people-focused research in general—is built on a foundation of research based on test subjects fitting that exact description. By doing so, the end result is that HCI-concerned designers are "creating technology that is optimized for WEIRD people" (p. 2426). The resulting designs may not be well-suited to or even usable by other populations and cultures.

A deep examination of this issue is far outside the scope of this paper, and indeed, an understanding of this issue and its ramifications in HCI research and elsewhere is in early stages. Tackling such an issue even in part may seem daunting, but as Henrich et al. (2010) note:

> "Recognizing the full extent of human diversity does not mean giving up on the quest to understand human nature. To the contrary, this recognition illuminates a journey into human nature that is more exciting, more complex, and ultimately more consequential than has previously been suspected." (p. 29)

That alone makes the effect of cultural background a worthy avenue of study. We surely cannot tackle the issue in its full scale, so for the purposes of this study, we will simply discuss some studies that strongly suggest there is, in fact, some evidence that cultural background has an observable effect on users' aesthetic sensibilities. Intuitively, we know that individuals will approach a given piece of art in very different ways, often due to some difference in personal background or inclination. It seems logical that this effect might also extend to aesthetics in web design, though we are only beginning to investigate and understanding the full implications of cultural background's effects on users' perceptions and tastes.

While not directly related to the concepts of clutter discussed above, which generally emphasized the organization of elements as a distinguishing factor, the choice and number of colors on a web page have been identified as some factors that determine the visual complexity of a given web page (Oliva et al., 2004; Rosenholtz et al., 2007). Cyr et al. (2009) specifically studied differences in which colors appealed to different cultures. They recruited their participants from students of three different countries of origin, focusing on those from Canadian, German, or Japanese background. Analysis of their results showed that perceptions of the color conditions had differences across cultural groups (Germans preferred the blue color scheme) but also some notable similarities (all groups disliked the yellow color scheme the most). Taken together, these findings suggest that cultural background can have an influence on what aspects of a website design people find appealing, while also suggesting some potential universalities in aesthetics the world over. Interestingly, participants of a German background did not explicitly refer to aesthetics having a positive or negative value when applied to websites.

The idea that different cultures may value aesthetics in and of themselves differently is one that hasn't been mentioned yet, but valuing or not valuing aesthetics may also have an influence on how people of that cultural background approach and process aesthetic aspects of an interface.

In addition to usability, visual design can affect trust and loyalty to a website, and these effects can differ depending on cultural background. Cyr (2014) notes that "in most studies when trust and culture are considered, the results are mixed or inconclusive… [h]owever, in one study, differences were found between collectivist (Chinese) and individualist (U.S.) cultures" (p. 50). This is not conclusive, but is suggestive of the wide-spanning effects of cultural background on how users perceive and interact with websites. From our aesthetics-focused standpoint, a more interesting finding is Cyr and Trevor-Smith's (2004) work in examining web site design across cultures. They found statistically significant differences among German, Japanese, and US web site designs, including differences in the use and preference for symbols and graphics, color choices, and navigational elements. This indirectly suggests that cultural background may affect aesthetic preferences, as web designers of a particular cultural background who are designing for users of that background would be apt to design according to the features considered most appealing by their intended users.

Reinecke and Gajos (2014) expand upon the above point by attempting to quantify differences in visual preferences around the world. Where Cyr et al. (2009) focused narrowly on three cultures, this study cast a far broader net and found more evidence toward the effect that cultural background has on users' aesthetic preferences. Using a first impression method, Reinecke and Gajos (2014) conducted an online test

where participants viewed a number of screenshots and ranked them on the site's level of 'visual appeal.' Each screenshot was displayed for no more than 500ms. They gathered data from nearly 40,000 volunteers from 179 countries. Their findings show that perceived visual complexity is a "strong predictor" of visual appeal, with designs that had low to medium visual complexity being the most appealing. In fact, they found that visual appeal was a more important predictor than colorfulness (p. 14). When describing the background of their research, Reinecke and Gajos (2014) mention that visual complexity is "[s]ometimes described with the negatively connoted term 'clutter'" (p. 12). For the purposes of their study, they seem to have conflated the two as 'visual complexity' and placed the concept in opposition to its lack, similar to how Stickel et al. (2010) treated the concept.

Reinecke and Gajos (2014) also found interesting findings upon examining the numerous demographic factors reported by their participants. For example, age affected preference for complexity—the older a participant, the more complex they preferred their websites. There are many possibilities for why this is the case, but the very existence of this effect points to the malleability of one's visual preferences in web design. Most relevant, however, are their findings involving country of origin. According to Reinecke and Gajos (2014), the results show "a significant interaction between a website's visual complexity and country" (p. 17). As an example of one such difference, they found that participants from Russia preferred the lowest level of visual complexity, versus participants from Serbia, who preferred websites with some of the highest visual complexity scores.

Tangentially, an interesting fact is that, in their demographic information, Reinecke and Gajos (2014) asked for current geographic location, rather than country of origin. Their choice of phrasing raises the question of whether one's aesthetic preferences may shift based on exposure to a different culture's aesthetics. If someone from the USA were to live a decade in a country with very different aesthetics, such as China, would they still prefer websites adhering to a USA aesthetic sensibility or would they begin to prefer a Chinese aesthetic? While this question is not addressed in Reinecke and Gajos' work, it raises an important point – the idea that aesthetics are a moving target. Googling the fashion of decades past is enough to demonstrate this idea. Aesthetic preferences shift, not just geographically, but chronologically as well. Therefore, rather than try to quantify the aesthetics of a given culture, comparative studies such as this one may be a good way to explore the effects of cultural background on aesthetics, as well as understanding *how* aesthetics differ between cultures.

## Hypotheses

The following hypotheses are based primarily on the work of Cheng and Nielsen (2016), who, in a study focused on comparing the reactions and interactions of Chinese versus American users of the same websites, found that "Chinese tend to complain less about complexity and prefer fairly high information density" (Introductory summary paragraph).

**H1:** Participants of Chinese background will rank websites of *all* clutter levels as *more* visually attractive than participants of American background.

**H2:** Participants of Chinese background will rank websites of *all* clutter levels as *more* organized in appearance.

**H3:** Participants of Chinese background will rank websites of the *high* clutter level as *less* cluttered in appearance.

**H4:** Participants of Chinese background will rank websites of *all* clutter levels has having *more* useful or important information.

## Methodology

The study was carried out through an online survey that was prepared and distributed through Qualtrics. Participants were UNC students of American and Chinese backgrounds. The survey asked questions about each participant's background and presented each participant with a series of 16 five-second impression tests of various website screenshots, including one training test that was not included for analysis.

### Sampling Plan

The survey was targeted at participants of American and Chinese backgrounds. As for our particular choice of cross-cultural examination, American- and Chinese-designed websites have very different approaches to design, making this a pairing with much potentially interesting data. Evidence for focus on this particular pairing may be seen in a recent cross-cultural study conducted by the Nielsen Norman Group, a prominent usability research and consulting agency. The online research report by Cheng and Nielsen (2016) points to the prominence of 'complexity' as a defining web design feature worthy of greater study, especially in a cross-cultural setting.

The criteria for participation were: enrollment at UNC; being over the age of 18; being of Chinese or American background; and being able to take the survey on something other than a mobile device. The minimum sample size goal was 20

participants in each cultural condition, for a total of 40 participants. There was no upper limit to sample size imposed.

Three recruitment methods were used to find participants. The primary recruitment method was distributing the survey via email. Student organizations serving Chinese international students (distinguished from Asian-American students) were identified and asked to forward a recruitment email explaining the survey's purpose and providing a link to the survey itself. The same thing was done for UNC academic departments that tend to have a higher-than-usual concentration of Chinese international students (as determined anecdotally). International students were not specifically targeted in the departmental emails. Emails were sent out to the entire relevant department, in hopes of capturing participants of both cultural background conditions.

The second recruitment method was via flyers. These were posted in public, high-traffic areas of the campus and included basic information about the survey, such as participation criteria and a link to the survey. The flyers did not otherwise attempt to target specific groups or individuals. The third recruitment method was through "snowball" recruitment. The closing page of the survey asked participants to forward the link to the survey to anyone they knew who might also be interested in participating. The second and third recruitment methods were meant to supplement the main recruitment method of email distribution.

**Survey Design and Materials**

The survey was created in Qualtrics and distributed electronically via email. The survey consisted of the following parts (full text in Appendix A):

1. *Consent form.* The first page of the survey reiterated the participation criteria and included the consent form information. Participants were prevented from proceeding until they acknowledged that they had read and understood the consent form.
2. *Demographic questions*. This section of the survey asked participants some brief questions about their enrollment status and, most importantly, used a pair of questions to capture information about each participant's cultural background. These questions are discussed in more detail below.
3. *Instructions.* The main portion of the survey was a series of 16 five-second impression tests, including a training test. Keeping in mind this format, an instruction page explaining the format and how to proceed was included, as well as a training set of screenshot and questions.
4. *Training set.* This set of web page screenshot and questions helped familiarize participants to the survey format. The data from this set was not analyzed.
5. *Test set.* This consists of the 15 five-second impression tests for which data was analyzed. Each test followed the same format: a web page screenshot was displayed for five seconds, followed by five questions. The order of screenshots was randomized for each participant. The choice of impression test format and the screenshot selection process will be discussed below.
6. *Closing.* A page thanking participants for their time and requesting that they forward the survey URL to any interested peers. This is for the snowball recruitment method described above.

### Capturing Cultural Background

The question of how to properly capture information about an individual's cultural background for the purposes of this study was complex to operationalize. Cultural background is an intensely personal topic and two individuals of similar upbringing may consider themselves to be of different cultures due to subjective senses of connection with their heritage(s). For the purposes of this study, we ultimately chose to follow the example set in Cyr (2008), as Cyr was engaged in similar work to that attempted in this paper—namely, that of understanding how cultural background affects reactions to website designs. Cyr notes that "[t]o ensure participants were 'of the culture' it was determined that each had lived in the country the majority of their lives and spoke the native language as their primary language" (p. 55).

In this study, we used two questions to capture both halves of Cyr's criteria—one to inquire as to where the participant had spent the majority of their life, and one to inquire as to the participant's native or primary language. For each question, the participant had the option of selecting USA, China, or writing in a third option. We felt that splitting Cyr's criteria into two halves would make it easier for participants to answer, rather than asking them to effectively consider two things in one question. The intention was to use the combination of answers to these two questions in order to identify each participant's cultural background. This framing of the cultural background did leave us open to the possibility of having invalid answers (if anyone selected the third write-in option) and of having to consider what to do with responses from participants who answered differently to each of the cultural background questions. The approach used to address these issues will be discussed in the Results section below.

### The 5-Second Test

The impression test format was chosen for this study because we were interested in participants' immediate reaction to each website's design. Therefore, in order to prevent participants from overthinking each website's design, each screenshot was displayed for no more than five seconds. Gronier (2016) tested the validity of the five-second test and found that it "captures the very first moments of human-computer interaction," further noting that "[t]hese few moments will affect the entire period of interaction" (p. 20). Since our area of concern in this study is how the first impressions of clutter in website design differ between participants of different cultural backgrounds, the five-second test was a good choice for that purpose. Furthermore, the short time span of the test meant that it would be possible to quickly gather first impressions of many websites in a single survey.

In order to capture the best first impression possible, Gronier (2016) recommends that the five-second test be applied to a static web page, a recommendation we followed. This notably means that the five-second test *cannot* assess the actual usability of websites, since there is no actual interaction involved. This test format can only capture information about the first reaction or impression. Any information captured about usability refers only to *perceived* usability, and how the aesthetic first impression of a given web page forms an impression about a website in the user's mind. The use of static screenshots also had many practical benefits, such as: ensuring that all participants viewed the exact same version of each of our chosen websites; that any moving parts, like carousel backgrounds, would not be present to distract participants as they took the

survey; and that each participant would view each web page for a precisely timed 5 seconds. This, of course, left the question of *which* screenshots to include in the survey.

### Screenshot Selection

The first step in the selection process was accessing the HTTP Archive[1], a website which tracks changes in web technology by analyzing the most high-traffic websites on the Internet. This list of high-traffic websites, which is publicly available on the HTTP Archive's website[2], is in turn provided by Alexa, a company that provides web traffic analytic tools. Alexa's traffic ranking data comes from two sources—user-installed browser extensions (it is not clear if these are installed for the specific purpose of allowing Alexa to collect data, or if these are functions bundled with extensions installed for other purposes) and scripts embedded in websites' code for direct tracking and certification of traffic metrics. Actual rankings are calculated using a proprietary method (Alexa). Unfortunately, Alexa only provides a partial list of the highest-traffic websites for free, so we had to use the HTTP Archive's list, which is available free up to the top 10,000 URLs. This was considered sufficient for our purposes, though we were not able to determine how old HTTP Archive's list of URLs was.

The final set of screenshots used in this study was pulled from the first 500 URLs generated by HTTP Archive's public list of the highest-traffic web sites. The list was saved to a Microsoft Word document for cleaning to remove duplicate websites (i.e. google.co.in and google.com are similar enough in design that it would be redundant to include both of them) and to potentially offensive websites (i.e. pornography). This

[1] http://httparchive.org
[2] http://httparchive.org/urls.php

cleaned list of URLs was then fed to WebShot[3], an open source tool that automatically takes screenshots of an uploaded list of URLs. Screenshots were taken in JPG format in order to use the JPG measure, where compressed file length predicts perceptions of image complexity (Donderi, 2005). While this program initially seemed like a time-saving measure to generate the screenshots, due to a number of technical issues, only the first 250 websites of the cleaned URL list were screenshotted using WebShot. Fortunately, there were enough valid screenshots in this set to show that 250 was sufficient to cover a wide range of file sizes. A new screenshot tool was found in qSnap, a Chrome browser extension[4], which produced higher quality JPEG screenshots than WebShot, but also required users to manually take screenshots. Due to time constraints, the WebShot screenshots were used as a rough guideline for sorting the screenshots into different categories of clutter level. The ones that were obviously faulty were removed, and the remainder (roughly 230) was sorted into three distinct groups, and from these three groups a final selection of 15 was selected. These fifteen were then re-screenshotted using qSnap, and the resulting file sizes were examined to ensure that the necessary category differentiations were retained.

However, the screenshot selection was not complete. Due to the varying completeness of the WebShot screenshots, they weren't an entirely accurate basis for the final set. The process of refining this final set of 15 was iterative. The initial WebShot screenshots were used as a rough guide to the relative sizes of the final screenshots. Likely candidate websites were selected in qSnap, and these screenshots were compared to the rest of the qSnap screenshots to see how well it fit with its intended category. This

[3] https://www.websitescreenshots.com/
[4] https://qsnapnet.com/

back-and-forth process was repeated several times, until the final 15 screenshots were deemed to be of three sufficiently differentiated categories of clutter as determined by the JPEG measure. While the WebShot screenshots were the rough guide to selecting the final set websites to be included in the survey, the final set of screenshots that were actually used in the survey were all taken in qSnap under the same settings.

The final set of screenshots falls into three "bands" of clutter level: high, medium, and low, with five screenshots per condition. They were all taken using qSnap in the Chrome web browser, on a widescreen laptop at 1366x768 resolution. The screenshots were from a mix of Western and Chinese websites, which was unintentional. The URLs and file size distribution are displayed below in Table 1:

*Table 1: Screenshot File Sizes and Classification*

| Clutter Level | High | Medium | Low |
|---|---|---|---|
| File Sizes (KB) | 1298 | 646 | 215 |
| | 1192 | 596 | 169 |
| | 982 | 498 | 163 |
| | 828 | 477 | 134 |
| | 782 | 430 | 107 |
| Average (KB) | 1016.4 | 529.4 | 157.6 |

**Data Collection**

At no point during the survey was identifying information like name, email, or student ID requested. Furthermore, Qualtrics' built-in anonymizing functions were also used to ensure participant confidentiality. The Anonymize Responses feature was used to ensure that not even IP addresses would be collected. To counter-act the possibility of a single participating taking the survey multiple times, Qualtrics' built-in 'prevent ballot box stuffing' feature was used. According to Qualtrics, there are ways this feature can be circumvented but given the low stakes nature of the survey, this was not considered to be

a significant risk. All data was collected within the Qualtrics platform. The data is presented in aggregate below.

## Results

### Cleaning & Coding

The final data set consisted of 63 responses. After removing all the incompletes—most of which were under 30% complete—the total number of responses was 33. While the initial survey included qualitative questions about each screenshot (see Appendix A), due to time constraints, these data will not be addressed in this paper. To prepare the data for analysis, two important issues needed addressing: identifying participants' cultural background and coding the Likert scales.

As noted above, two questions asked participants for information that was used to categorize them by cultural background: one about the country in which they had lived the majority of their lives and one about their primary language. This two-question format was chosen for a number of practical reasons, but it did open up the possibility of encountering cases where, for example, someone chose China as their country and English as their primary language. Fortuitously, no such cases were present in the full responses to the survey. Since each participant's language matched their chosen country, the country variable was used as shorthand for the participants' cultural background in the data analysis. As for the coding of the Likert scales, each response was mapped onto a number for the analysis, as shown in Table 2.

*Table 2: Scales and Codes*

| Code | Scale 1 | Scale 2 |
|---|---|---|
| 1 | Strongly disagree | Far too little |
| 2 | Disagree | Moderately too little |
| 3 | Somewhat disagree | Slightly too little |
| 4 | Neither agree nor disagree | Neither too much nor too little |
| 5 | Somewhat agree | Slightly too much |
| 6 | Agree | Moderately too much |
| 7 | Strongly agree | Far too much |

All of the following statistical analyses were conducted in R.

**Participant Characteristics**

The final participant pool was 33 full responses, with 24 from the USA cultural background and 9 from the Chinese cultural background. As noted above, the anticipated response was at least 20 per cultural condition, so overall, these numbers fall short of what was hoped for. All statistical analyses were performed over all data despite the unequal group sizes. However, these two groups were notably similar in some respects.

*Table 3: Participant Ages*

| Statistic | China | USA | Total |
|---|---|---|---|
| Mean Age | 26.00 | 25.21 | 25.42 |
| Min. Age | 19 | 19 | 19 |
| Max. Age | 39 | 40 | 40 |

*Table 4: Participant Education Levels*

| Level | China | USA | Total |
|---|---|---|---|
| Undergraduate | 1 | 7 | 8 |
| Master's student | 5 | 13 | 18 |
| Doctoral Candidate | 2 | 4 | 6 |
| Other (Please specify.) | 1 | 0 | 1 (didn't specify) |

*Table 5: Participant Desktop/Laptop Web Browsing Frequency*

| Frequency | China | USA | Total |
|---|---|---|---|
| At least once a day | 9 | 20 | 29 |
| 2-3 days in a week | 0 | 2 | 2 |
| Once every two weeks | 0 | 1 | 1 |
| Once a month | 0 | 0 | 0 |
| Rarely | 0 | 1 | 1 |

As Table 3 shows, the two groups have very similar mean ages and age ranges. Table 4 further shows that the two groups had roughly the same distribution over enrollment levels, with the majority of respondents in both groups being Master's students (the survey didn't ask for major or program information). Table 5 shows that the groups were also similar in their self-reported Internet usage. The question capturing this information asked participants how often they used a laptop or desktop to browse the Internet. The vast majority reported using a laptop or desktop computer to browse the Internet at least once a day, with a few participants indicating less frequent use in the larger USA background group. From this we can assume that most, if not all, participants were reasonably familiar with Web aesthetics for desktop browsing.

**Exploratory Analysis**

For the initial exploratory analysis, the rating data were grouped according to the three clutter levels represented in the screenshots: High, Medium, and Low clutter. Additionally, the data was analyzed according to each of the four Likert questions represented in the survey. These analyses are represented below using the following shorthand for the Likert questions (refer to Table 2 above for the full set of values present on the two scales used.):

- *VisAttr*: Scale 1, Agree/Disagree, "This web page is visually attractive."
- *Org*: Scale 1, Agree/Disagree, "This web page is well organized."
- *Clutter*: Scale 1, Agree/Disagree, "This web page is cluttered."
- *InfoUse*: Scale 2, Too Much/Too Little, "How much of this page had **useful or important** information?" [emphasis in original, see Appendix A]

Tables 6-8 below present some descriptive statistics of the Likert question responses.

*Table 6: Mean Responses to Clutter Level = High*

| Likert Q | Statistic | China | USA | Total |
|---|---|---|---|---|
| VisAttr | Mean | 3.96 | 3.63 | 3.72 |
| | SD | 2.07 | 2.12 | 2.11 |
| | Median | 4.00 | 3.00 | 3.00 |
| Org | Mean | 4.67 | 4.24 | 4.35 |
| | SD | 1.54 | 1.79 | 1.73 |
| | Median | 5.00 | 5.00 | 5.00 |
| Clutter | Mean | 4.73 | 5.03 | 4.95 |
| | SD | 1.60 | 1.91 | 1.83 |
| | Median | 5.00 | 6.00 | 5.00 |
| InfoUse | Mean | 4.73 | 4.66 | 4.68 |
| | SD | 1.57 | 1.67 | 1.64 |
| | Median | 5.00 | 4.00 | 4.00 |

*Table 7: Mean Responses to Clutter Level = Medium*

| Likert Q | Statistic | China | USA | Total |
|---|---|---|---|---|
| VisAttr | Mean | 5.00 | 4.83 | 4.87 |
| | SD | 1.35 | 1.90 | 1.76 |
| | Median | 5.00 | 5.00 | 5.00 |
| Org | Mean | 5.18 | 4.74 | 4.86 |
| | SD | 1.40 | 1.71 | 1.64 |
| | Median | 5.00 | 5.00 | 5.00 |
| Clutter | Mean | 3.76 | 3.64 | 3.67 |
| | SD | 1.58 | 1.94 | 1.85 |
| | Median | 4.00 | 3.50 | 4.00 |
| InfoUse | Mean | 4.09 | 3.73 | 3.83 |
| | SD | 1.16 | 1.42 | 1.36 |
| | Median | 4.00 | 4.00 | 4.00 |

*Table 8: Mean Responses to Clutter Level = Low*

| Likert Q | Statistic | China | USA | Total |
|---|---|---|---|---|
| VisAttr | Mean | 5.09 | 4.61 | 4.74 |
| | SD | 1.67 | 1.86 | 1.81 |
| | Median | 5.00 | 5.00 | 5.00 |
| Org | Mean | 5.25 | 5.37 | 5.34 |
| | SD | 1.30 | 1.28 | 1.28 |
| | Median | 5.00 | 6.00 | 5.00 |
| Clutter | Mean | 3.48 | 2.45 | 2.72 |
| | SD | 1.69 | 1.44 | 1.57 |
| | Median | 4.00 | 2.00 | 2.00 |
| InfoUse | Mean | 4.00 | 3.46 | 3.61 |
| | SD | 1.16 | 1.07 | 1.12 |
| | Median | 4.00 | 4.00 | 4.00 |

**ANOVA & Post Hoc Tests**

To analyze the ratings data, we used a two-way factorial ANOVA[5] test. Cultural Background (Levels: China, USA) and Clutter Level (Levels: High, Medium, Low) were treated as independent variables. Ratings were treated as the dependent variable. We conducted a total of four ANOVA tests, one per Likert question in our set. The results of the ANOVA tests are presented below in Tables 9-12:

*Table 9: ANOVA Results-VisAttr*

| | df | Sum Sq. | Mean Sq. | F-value | Pr(>F) | Signif. |
|---|---|---|---|---|---|---|
| Culture | 1 | 10.2 | 10.15 | 2.814 | 0.0941 | |
| ClutterLvl | 2 | 131.1 | 65.56 | 18.173 | $2.45x10^{-8}$ | *** |
| Culure:ClutterLvl | 2 | 1.5 | 0.74 | 0.206 | 0.8143 | |
| Residuals | 487 | 1757.0 | 3.61 | | | |

Signif. codes: * indicates p<0.05; ** p<0.01; *** p<0.001
2 observations deleted due to missing values.

[5] Without modifications, R is only able to run the Type I ANOVA test. Due to time constraints I wasn't able to implement the Type III ANOVA in R, which may have affected the validity of the results.

*Table 10: ANOVA Results-Org*

| | df | Sum Sq. | Mean Sq. | F-value | Pr(>F) | Signif. |
|---|---|---|---|---|---|---|
| Culture | 1 | 6.0 | 5.98 | 2.458 | 0.118 | |
| ClutterLvl | 2 | 79.3 | 39.66 | 16.295 | $1.41 \times 10^{-7}$ | *** |
| Culure:ClutterLvl | 2 | 6.6 | 3.30 | 1.356 | 0.259 | |
| Residuals | 486 | 1183.0 | 2.43 | | | |

Signif. codes: * indicates p<0.05; ** p<0.01; *** p<0.001
3 observations deleted due to missing values.

*Table 11: ANOVA Results-Clutter*

| | df | Sum Sq. | Mean Sq. | F-value | Pr(>F) | Signif. |
|---|---|---|---|---|---|---|
| Culture | 1 | 7.8 | 7.77 | 2.577 | 0.10909 | |
| ClutterLvl | 2 | 409.9 | 204.93 | 67.939 | $<2 \times 10^{-16}$ | *** |
| Culure:ClutterLvl | 2 | 30.1 | 15.04 | 4.985 | 0.00719 | *** |
| Residuals | 487 | 1468.9 | 3.02 | | | |

Signif. codes: * indicates p<0.05; ** p<0.01; *** p<0.001
2 observations deleted due to missing values.

*Table 12: ANOVA Results-InfoUse*

| | df | Sum Sq. | Mean Sq. | F-value | Pr(>F) | Signif. |
|---|---|---|---|---|---|---|
| Culture | 1 | 10.4 | 10.39 | 5.421 | 0.0203 | * |
| ClutterLvl | 2 | 104.5 | 52.25 | 27.262 | $6.05 \times 10^{-12}$ | *** |
| Culure:ClutterLvl | 2 | 3.6 | 1.79 | 0.932 | 0.3943 | |
| Residuals | 483 | 925.8 | 1.92 | | | |

Signif. codes: * indicates p<0.05; ** p<0.01; *** p<0.001
6 observations deleted due to missing values.

As Tables 9-12 show, all four of the Likert questions resulted in some statistically significant effects of the Culture or Clutter Level variables at the $\alpha = 0.05$ level. Table 9 shows that, for the VisAttr question, the effect of Culture only approached significance. This is surprising, as previous reading suggested that there would be a significant difference in reactions to display clutter between cultures, especially the particular cultural pairing studied here. However, the effect of Clutter Level was definitely significant. Table 10 shows that, for the Org question, Clutter Level had a significant effect on the results, which was expected. However, no significant effect of Culture was detected, so no support was found for hypothesis H2. Table 11 shows that, for the Clutter

question, there was a significant effect of Clutter Level, which is as expected. There was also an interaction effect of Culture and Clutter Level. Finally, Table 12 shows that, for the InfoUse question, there was a significant main effect of Culture and a significant effect of Clutter Level.

Of note is that, for each ANOVA, some observations had to be ignored due to non-response throughout the survey. This was likely because—except for the consent form question on the first page—no question on the survey required an answer in order for the respondent to proceed. It is possible that these nonresponses came from participants forgetting to answer or choosing not to due to indecision or fatigue. While these nonresponses are scattered throughout the final data set, overall they weren't considered numerous enough to endanger the analysis. The VisAttr ANOVA was missing one observation per Culture group; Org was missing one from China and two from USA; Clutter was missing one per group; and InfoUse was missing one from China and five from USA.

Figures 1-3 below present some further comparisons between the mean responses to each question.

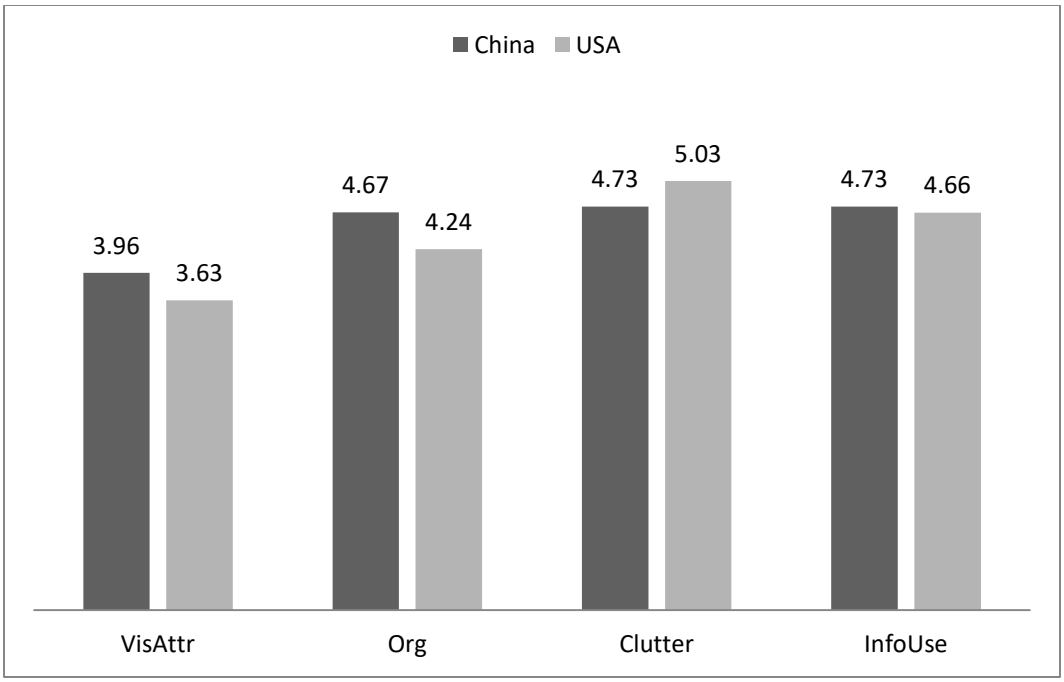


*Figure 1: High Clutter Mean Response Comparison. This graph presents the mean responses to questions about screenshots in the High clutter level category.*

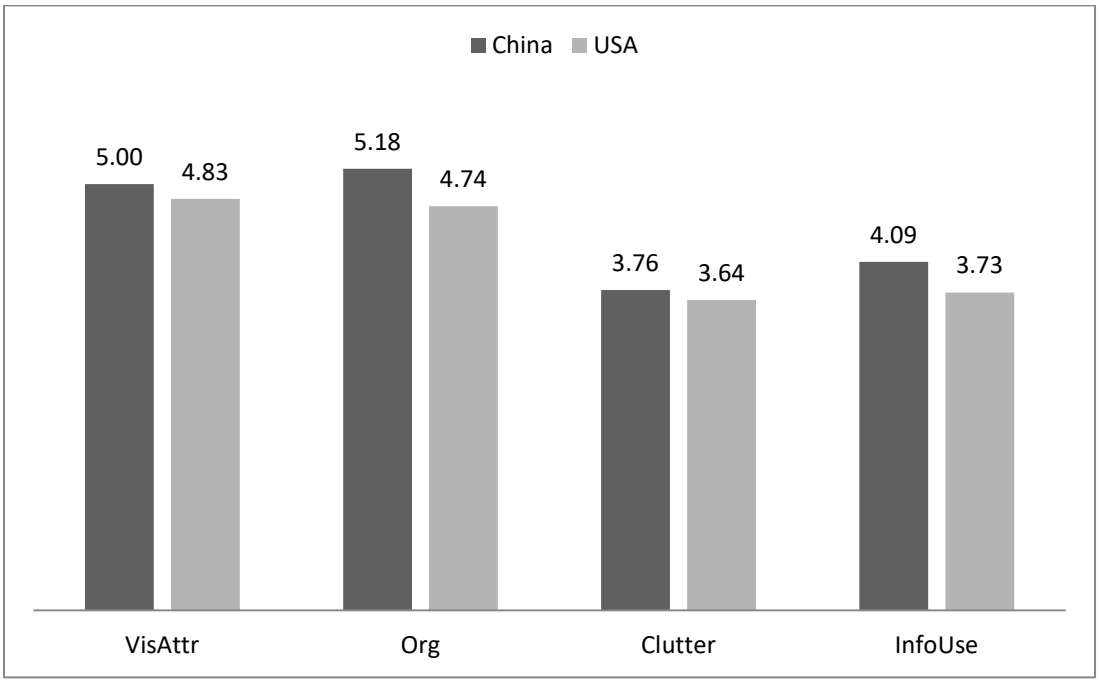


*Figure 2: Medium Clutter Mean Response Comparison. This graph presents the mean responses to questions about screenshots in the Medium clutter level category*

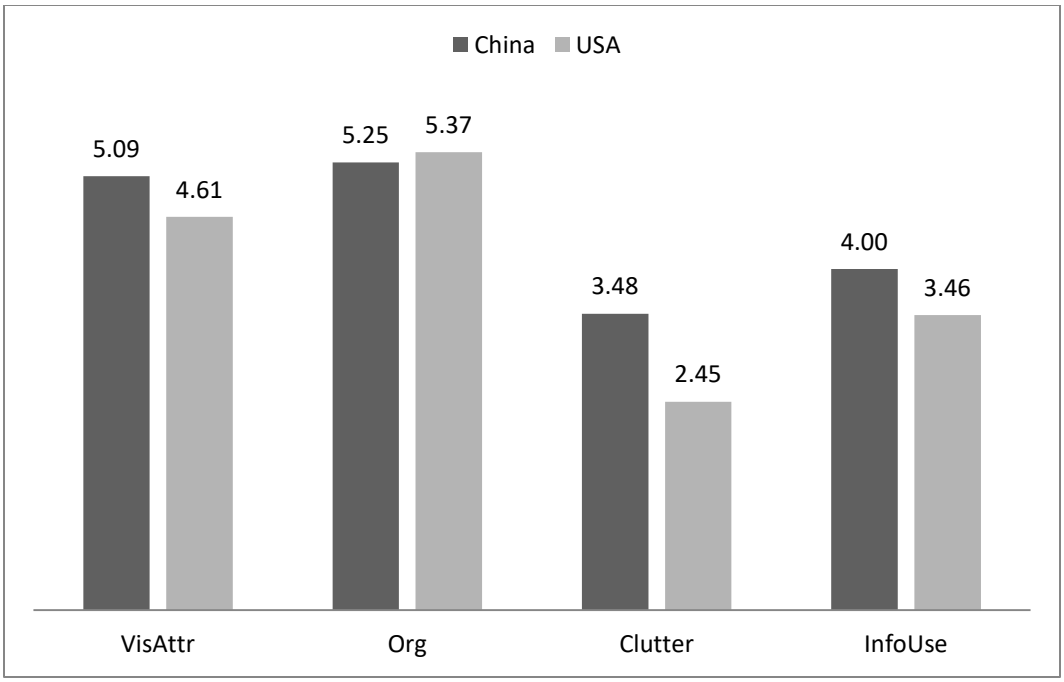


*Figure 3: Low Clutter Mean Response Comparison. This graph presents the mean responses to questions about screenshots in the Low clutter level category*

These figures bear out the ANOVA results. There seems to be an overall trend of the China participants giving slightly higher ratings than the USA participants. However, Figure 3 is notable because it has the largest visible difference between means for the responses to the Clutter question. This further supports the results shown in Table 11, which show the existence of effects from Clutter Level and the interaction of Culture and Clutter Level.

The next step was to conduct post-hoc analyses in order to pinpoint the exact variables where this differentiation was occurring. For this, we chose to use Tukey's 'Honest Significant Difference' test. One of these analyses was run for each of the four ANOVAs. The relevant results are presented and discussed below.

*Table 13: Post-hoc Analysis-VisAttr*

| Variable | Comparison | Difference | 95% CI lower bound | 95% CI upper bound | p-value |
|---|---|---|---|---|---|
| Culture | USA-China | -0.32 | -0.70 | 0.06 | 0.09 |
| ClutterLvl | High-Med | 1.15 | 0.66 | 1.64 | 0.00 |
| | Med-Low | 0.13 | -0.36 | 0.62 | 0.81 |
| | High-Low | 1.02 | 0.53 | 1.52 | 0.00 |

Non-significant results in grey.

For the effect of Clutter Level on VisAttr, the post-hoc comparisons showed significant differences between the High and Medium levels and the High and Low levels (Table 13). Based on these results, we can neither fully reject nor fully support hypothesis H1 (Participants of Chinese background will rank websites of *all* clutter levels as *more* visually attractive than participants of American background). However, H1 may merit further examination since the main effect of Culture approached significance (p=0.09).

*Table 14: Post-hoc Analysis-Org*

| Variable | Comparison | Difference | 95% CI lower bound | 95% CI upper bound | p-value |
|---|---|---|---|---|---|
| ClutterLvl | High-Med | 0.51 | 0.10 | 0.91 | 0.01 |
| | Med-Low | -0.48 | -0.88 | -0.07 | 0.02 |
| | High-Low | 0.99 | 0.58 | 1.39 | 0.00 |

In terms of the effect of Clutter Level on the Organization question ratings, all possible pairings of Clutter Level were significantly different from each other (Table 14). This shows that users had strong distinctions between clutter levels when judging the web sites' visual organization. However, since the main effect of Culture on Organization ratings did not reach significance, we did not find support for H2 (Participants of Chinese background will rank websites of *all* clutter levels as *more* organized in appearance.).

*Table 15: Post-hoc Analysis-Clutter*

| Variable | Comparison | Diff. | 95% CI lower bound | 95% CI upper bound | p-value |
|---|---|---|---|---|---|
| ClutterLvl | High-Med | -1.28 | -1.73 | -0.83 | 0.0e+00 |
| | Med-Low | 0.95 | 0.50 | 1.40 | 3.2e-06 |
| | High-Low | -2.23 | -2.68 | -1.78 | 0.0e+00 |
| Culture& | USA:High-China:High | 0.30 | -0.57 | 1.17 | 0.92 |
| ClutterLvl | China:Low-China:High | -1.26 | -2.31 | -0.20 | 0.01 |
| | USA:Low-China:High | -2.29 | -3.16 | -1.42 | 0.00 |
| | China:Med-China:High | -0.98 | -2.03 | -.07 | 0.08 |
| | USA:Med-China:High | -1.09 | -1.96 | -0.22 | 0.01 |
| | China:Low-USA:High | -1.56 | -2.43 | -0.68 | 0.00 |
| | USA:Low-USA:High | -2.59 | -3.23 | -1.95 | 0.00 |
| | China:Med-USA:High | -1.28 | -2.15 | 0.41 | 0.00 |
| | USA:Med-USA:High | -1.39 | -2.03 | -0.75 | 0.00 |
| | USA:Low-China:Low | -1.03 | -1.91 | -0.16 | 0.01 |
| | China:Med-China:Low | -.28 | -0.78 | 1.33 | 0.97 |
| | USA:Med-China:Low | 0.16 | -0.71 | 1.04 | 1.00 |
| | China:Med-USA:Low | 1.31 | 0.44 | 2.18 | 0.00 |
| | USA:Med-USA:Low | 1.20 | 0.55 | 1.84 | 0.00 |
| | USA:Med-China:Med | -0.11 | -0.98 | 0.75 | 1.00 |

Non-significant results in grey.

In terms of the effect of Clutter Level on the Clutter question ratings, all combinations of Clutter Levels were significantly different from each other (Table 15). Again, this suggests that participants' impressions of how cluttered the web sites were matched well to the categories of JPEG measure used to differentiate between levels of Clutter in the web sites. As these results were in a statement that asked for agreement or disagreement with whether or not a given web site gave the impression of being cluttered, this suggests that the JPEG measure does in fact match well to subjective judgments of what is 'cluttered' in a web site (not just to what is 'visually complex').

The Culture x Clutter Level combinations are numerous. However, the pairs where the values of *both* Culture and Clutter Level vary are not of interest. This is because, for our purposes, we are most interested in a) when different cultures have a

significantly different reaction to the same amount of clutter or b) when the same culture has a significantly different reaction to different amounts of clutter. We consider the following comparisons of interest:

- Culture = China; Clutter Level: Low vs. High
- Culture = USA; Clutter Level: High vs. Med, Med vs. Low, High vs. Low
- Culture = USA vs. China; Clutter Level: Low

There are two things of interest visible in these results. First, both the China and USA groups show some significant difference in reaction based on Clutter Level. Even more interesting is the significant effect of Culture at the Low Clutter Level. Based on these results, we did not find evidence to support H3. However, we did find an interaction effect that indicated that participants in the China background group rated the Low clutter screenshots higher than the participants in the USA background group. This is supported by the difference in mean ratings seen between the Clutter bars in Figure 3. As Table 2's scales indicate, the higher rating indicates a higher level of agreement with the statement "This web page is cluttered."

*Table 16: Post-hoc Analysis-InfoUse*

| Variable | Comparison | Difference | 95% CI lower bound | 95% CI upper bound | p-value |
|---|---|---|---|---|---|
| Culture | USA-China | -0.33 | -0.60 | -0.05 | 0.02 |
| ClutterLvl | High-Med | -0.85 | -1.21 | -0.49 | 0.00 |
| | Med-Low | 0.22 | -0.14 | 0.58 | 0.33 |
| | High-Low | -1.07 | -1.43 | -0.71 | 0.00 |

Non-significant results in grey.

In terms of InfoUse, Table 16 shows two findings of interest. First, there is the significant difference in the InfoUse ratings between the High and Medium clutter levels and between the High and Low clutter levels. Even more interesting, there is a significant

main effect of Culture on the ratings related to the amount of useful information in the screenshots. Based on these results, there is some partial support for H4 (Participants of Chinese background will rank websites of *all* clutter levels has having *more* useful or important information.). Reviewing Figures 1-3 shows that, for the Medium and Low Clutter Levels, the China group had noticeably higher ratings of how much useful information was present on the screenshots. However, for the High Clutter Level, the difference in rating between the two groups was much smaller.

## Conclusion

In this study, we found effects of Culture, but did not detect the effects for all the variables that we investigated. First, Culture had a main effect on InfoUse. These were Likert ratings in response to the statement "How much of this page had **useful or important** information?" It seems like participants with the China background tended overall to rate the screenshots has having higher amounts of useful or important information. There was also a significant effect of the interaction of Culture and Clutter Level for the question about clutter.

Our analyses found effects of Clutter Level on all four of the main variables that we investigated. The Clutter Levels were established based on the JPEG measure, and the existence of main effects from Clutter Level supports the idea of a relationship between the JPEG measure's measurement of clutter and subjective assessments of clutter. Larger JPEG size does indeed appear to be related to viewer perception of increased clutter. This supports Donderi (2005)'s finding that the JPEG measure correlated to subjective perceptions of visual complexity. However, it's notable that in most cases, these effects seemed more distinct for some questions than others. For VisAttr and InfoUse, the effect

only reached significance when comparing High and Medium levels and the High and Low levels. However, for Org and Clutter, the effect was significant for High and Medium, High and Low, *and* Medium and Low, covering all possible combinations. It may be that the JPEG measure, as expressed through the Clutter Levels, is more closely related to viewer perceptions of the latter characteristics (the appearance of organization and whether or not it is perceived to be cluttered) than the former (the appearance of visual attractiveness and perceived amount of useful information).

The lack of significant effect between the Medium and Low Clutter Levels for VisAttr and InfoUse may be due to a number of factors. One possibility is that, when selecting the screenshots, the Medium and Low representatives weren't differentiated enough. However, as the Org and Clutter questions showed significant effects for that specific pairwise comparison, it's not clear whether or not a lack of distinction between Medium and Low when selecting the initial screenshots was actually present. Another possibility is that, when it came to visual attractiveness and the perceived amount of useful information, the participants' judgment was not along a scale but more along a "yes/no" approach—either the screenshot in question possessed the characteristic, or it did not.

For Org and Clutter, there was a significant effect between all pairwise comparisons of Clutter Level. This supports the idea that organization may be a distinguishing factor between clutter and visual complexity, a finding that has popped up repeatedly in the research literature (Oliva et al., 2004; Stickel et al., 2010; Moacdieh and Sarter, 2014; Rosenholtz et al., 2007). The Clutter results reveal an additional result of interest, showing a significant interaction effect between Culture and Clutter Level.

Looking at Figure 3, we see that the mean rating for the China group was 3.48 (between "Somewhat disagree" and "Neither disagree nor agree") whereas the mean rating of the USA group was 2.45 (between "Disagree" and "Somewhat disagree"). The USA group more strongly identified the Low screenshots as not being cluttered.

Another interesting aspect of the interaction effect is that for the China group, the only difference between Clutter Levels that reached significance was between the High and Low levels, the strongest possible distance between levels of clutter. For the USA group, all possible pairings of Clutter Levels resulted in significant effects. The JPEG measure was our basis for determining these levels, and as noted earlier, it has been shown to predictive subjective judgments of visual complexity (Donderi, 2005). Given our results, future work could explore whether the JPEG measure more closely matches the distinctions about clutter perceptions of one cultural group versus another. The findings here only weakly suggest this possibility, due to the small size of the China group, but this could be an avenue for future work in this area.

A great deal more work needs to be done in cross-cultural reactions to clutter and web aesthetics in general before any conclusive statement can be reached. Rather than shy away from these potentially rich areas of research and understanding, we should embrace them.

## References

Alexa Internet – About Us. (n.d.) Retrieved from http://www.alexa.com/about. Last visited 4/30/2017.

Cheng, Y., & Nielsen, J. (2016). Are Chinese websites too complex? Retrieved from https://www.nngroup.com/articles/china-website-complexity/

Cyr, D. (2008). Modeling web site design across cultures: Relationships to trust, satisfaction, and e-loyalty. *Journal of Management Information Systems, 24*(4), 47-72. http://dx.doi.org/10.2753/MIS0742-1222240402

Cyr, D., Head, M., & Larios, H. (2010). Colour appeal in website design with and across cultures: A multi-method evaluation. *International Journal of Human-Computer Studies, 68*(1), 1-21. http://dx.doi.org/10.1016/j.ijhcs.2009.08.005

Cyr, D., & Trevor-Smith, H. (2004). Localization of Web design: An empirical comparison of German, Japanese, and United States web site characteristics. *Journal of the American Society for Information Science and Technology, 53*(13), 1199-1208. DOI: 10.1002/asi.20075

Donderi, D. C., & McFadden, S. (2005). Compressed file length predicts search time and errors on visual displays. *Displays, 26,* 71-78. http://dx.doi.org/10.1016/j.displa.2005.02.002

Donderi, D. C. (2006). An information theory analysis of visual complexity and dissimilarity. *Perception, 35,* 823-835. DOI: 10.1068/p5249

Gronier, G. (2016). Measuring the first impression: Testing the validity of the 5 second test. *Journal of Usability Studies, 12*(1), 8-25. Retrieved from http://uxpajournal.org/measuring-testing-validity-5-second-test/

Harper, S., Michailidou, E., & Stevens, R. (2009). Toward a definition of visual complexity as an implicit measure of cognitive load. *ACM Transactions on Applied Perception (TAP), 6*(2), 1-18. DOI: 10.1145/1498700.1498704

Henrich, J., Helne, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature, 466*(7302), 29. Retrieved from http://libproxy.lib.unc.edu/login?url=http://search.proquest.com/docview/610444800?accountid=14244

Lindgaard, G., Dudek, C., Sen, D., Sumegi, L., & Noonan, P. (2011). An exploration of relations between visual appeal, trustworthiness and perceived usability of homepages. *ACM Transactions on Computer-Human Interaction, 18*(1), 1-30. DOI: 10.1145/1959022.1959023

Lindgaard, G., Fernandes, G., Dudek, C., & Brown, J. (2006). Attention web designers: You have 50 milliseconds to make a good first impression! *Behaviour & Information Technology, 25*(2), 115-126. http://dx.doi.org/10.1080/01449290500330448

Moacdieh, N., & Sarter, N. (2014). Display clutter: A review of definitions and measurement techniques. *Human Factors, 57*(1), 61-100. DOI: 10.1177/0018720814541145

Oliva, A., Mack, M. L., Shrestha, M., & Peeper, A. (2004). Identifying the perceptual dimensions of visual complexity of scenes. *Proceedings of the Annual Meeting of the Cognitive Science Society, 26*. http://escholarship.org/uc/item/17s4h6w8

Reinecke, K., & Gajos, K. Z. (2014). Quantifying visual preferences around the world. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 11-20. DOI: 10.1145/2556288.2557052

Rosenholtz, R., Li, Y., & Nakano, L. (2007). Measuring visual clutter. *Journal of Vision, 7*(2), 1–22. DOI: 10.1167/7.2.17

Stickel, C., Ebner, M., & Holzinger, A. (2010) The XAOS metric – Understanding visual complexity as measure of usability. In Leitner, G., Hitz, M., & Holzinger, A. (Eds.), *HCI in Work and Learning, Life and Leisure* (278-290). Berlin, Heidelberg: Springer. DOI: 10.1007/978-3-642-16607-5_18

Sturm, C., Oh, A., Linxen, S., Abdelnour-Nocera, J., Dray, S., & Reinecke, K. (2015). How WEIRD is HCI? Extending HCI principles to other countries and cultures. *CHI'15 Extended Abstracts*. http://dx.doi.org/10.1145/2702613.2702656

Tractinsky, N., Katz, A. S., & Ikar, D. (2000). What is beautiful is usable. *Interacting with Computers, 13*(2). 127-145. https://doi-org.libproxy.lib.unc.edu/10.1016/S0953-5438(00)00031-X

## Appendix A: Survey Text

### Page 1 – Consent Form

Welcome!

This survey is for a research study on how different cultural backgrounds may have an effect on how people react to website design. During this survey, you'll be asked to view a series of screenshots for 5 seconds each. You'll be asked 5 questions about each screenshot. The entire survey will take about 15 minutes to complete. You may stop taking the survey and quit at any time.

Participation in this study is completely voluntary and anonymous. Your responses will be used for research purposes only. Any data from this survey will be presented anonymously or in aggregate in order to protect your privacy.

Please only participate if you:

- Are a UNC student
- Are over 18 years of age
- Mostly grew up in the USA or in China
- Are reading this on a laptop or desktop computer (This survey is not designed for phones or tablets.)

This study has been reviewed and approved by the UNC-CH Institutional Review Board, a committee that works to protect your rights and welfare as a research study volunteer. If you have any questions or comments about this study, contact the IRB at IRB_subjects@unc.edu. This study was reviewed and approved as study number #17-0337.

By selecting "I Agree" below, you're stating that you

- understand the information on this page,
- fulfill the participation requirements, and
- agree to participate in this study.

[__radio button__] I Agree[6]

[6] This question uses Qualtrics' built-in validation functions to ensure that participants can't continue to the rest of the survey without selecting yes. This was the *only* question that with a required answer. All other questions were optional.

**Page 2 – Background Info**

Please answer the following questions about your background.

How old are you?

[__text entry__]

Which of these options describes you best? (Radio buttons.)

- Undergraduate student
- Master's student
- Doctoral candidate
- Other (Please specify.) [__text entry__]

Which country have you lived in for the **majority**[7] of your life? This is defined as the country where you have lived for the **greatest number** of years.

- USA
- China
- Other (Please specify.) [__text entry__]

What is your **primary or native** language? This is defined as the language you **learned first** or feel **most comfortable** using in daily life.

- English
- Chinese, any dialect (Mandarin, Cantonese, etc.)
- Other (Please specify.) [__text entry__]

How often do you use a laptop or desktop computer to surf the Internet? (Radio buttons.)

- At least once a day
- 2-3 days in a week
- Once every two weeks
- Once a month
- Rarely

[7] The text styling here and throughout all later questions is included as it appeared in the survey.

**Page 3 - Instructions**

For the next part of the survey, we will present you with a series of screenshots from different websites. Each screenshot will appear for only 5 seconds. Please focus on the screenshot while it's displayed.

Each screenshot will be followed by 5 questions. What we're interested in for this study is your first impression, so don't worry about picking the "right" answer. Answer in whatever way seems most right to you.

A total of 16 screenshots will be displayed. The very first screenshot is a training screenshot to help you get used to this format.

Ready? Click the ">>" button to continue.[8]

**Page 4 – Test Screenshot**

[This page displays a screenshot of the UNC home page (http://www.unc.edu/).]

**Page 5 – Test Questions**

List **1 to 3 words** describing the web page you just saw. [__text entry__]

These statements are about your **first impression** of the screenshot you just saw. [Likert matrix of radio buttons.]

Statements:

- This web page is visually attractive.
- This web page is well organized.
- This web page is cluttered

Scale:

- Strongly agree
- Agree
- Somewhat agree
- Neither agree nor disagree
- Somewhat disagree
- Disagree
- Strongly disagree

[8] The default "next" button in Qualtrics has ">>" on it and appeared on this page of the survey in the lower right corner of the screen.

This question is about the **kind of information** that was in the screenshot you just saw.

Statement:
- How much of this page had **useful or important** information?

Scale:
- Far too much
- Moderately too much
- Slightly too much
- Neither too much nor too little
- Slightly too little
- Moderately too little
- Far too little

**Pages 6 to 35 – Screenshots and Questions 1 to 15**

[The next 30 pages repeated the format of pages 4-5 using the 15 actual screenshots. URLs of the screenshots used are provided below.

High
- huaban.com
- bing.com
- sohu.com
- xinhuanet.com
- seattle.craigslist.org

Medium:
- wellsfargo.com
- daumn.net
- shutterstock.com
- stackoverflow.com
- alipay.com

Low
- thepiratebay.org
- zhihu.com
- apple.com
- openload.co
- baidu.com

The screenshots were presented in randomized order for each participant.]

**End of Survey – Closing and Thank You**

## Thank you for taking this survey! Your response has been recorded.

## If you know anyone you know who might be interested in taking this survey, please forward them this link: bit.ly/ClutterCulture.

If you have any questions or comments, you may contact the principle investigator, Stephanie Hsieh, at syfh@live.unc.edu.

As a reminder, you may also (anonymously) contact the University of North Carolina at Chapel Hill's Institutional Review Board with your concerns or comments about this study. You may contact the IRB at IRB_subjects@unc.edu. This study was reviewed and approved as study number #17-0337.