Using Biometrics to Evaluate Visual Design

Andrew Coles, Dixie Hamilton, Peace Iyiewuare
School of Information, The University of Texas at Austin, Austin, TX
andrewpcoles@gmail.com, dixie.hamilton@gmail.com, piyiewuare123@gmail.com

Abstract
Visual design is a critical aspect of any web page or user interface, and its impact on a user's experience has been studied extensively. Research has shown a positive correlation between a user's perceived usability and a user's assessment of visual design. Additionally, perceived web quality, which encompasses visual design, has a positive relationship with both initial and continued consumer purchase intention. However, visual design is often assessed using self-report scale, which are vulnerable to a few pitfalls. Because self-report questionnaires are often reliant on introspection and honesty, it is difficult to confidently rely on self-report questionnaires to make important decisions. This study aims to ensure the validity of a visual design assessment instrument (Visual Aesthetics of Websites Inventory: Short version) by examining its relationship with biometric (variables), like galvanic skin response, pupillometry, and fixation information. Our study looked at participants assessment of a webpage's visual design, and compared it to their biometric responses while viewing the webpage. Overall, we found that both average fixation duration and pupil dilation differed when participants viewed web pages with lower visual design ratings compared to web pages with a higher visual design rating.

Keywords
usability, visual design, websites, eye tracking, pupillometry, self-report, VisAWI

Introduction
A vast amount of research has been conducted regarding the importance of visual design, and its role as a mediator of user's experience when browsing a site or interacting with an interface. In the literature, visual design is one aspect of website quality. Jones and Kim (2010) define website quality as "the perceived quality of a retail website that involves a [user's] perceptions of the retailer's website and comprises consumer reactions towards such attributes as information, entertainment/enjoyment, usability, transaction capabilities, and design aesthetics." Jones and Kim (2010) examined the impact web quality and retail brand trust has on purchase intentions. Additional research examining e-commerce sites has shown web quality has an impact on both initial and continued purchase intention (Bock & Vathanophas, 2008), as well as consumer satisfaction (Lin, 2010). Moreso, research on the relationship between visual design and perceived usability (Lindgaard, Pilgrim, Stojmenovic, 2014) has revealed a positive correlation between the two. As users' ratings of visual quality increase, their ratings of perceived usability follows a similar trend. Although this research spans various domains, the reliance on self-report measures to gauge concepts like visual design and web quality is prevalent throughout much of the literature.

Although some self-report scales are validated within the literature, there are still issues with the use of self-report questionnaires. One is the reliance on the honesty of the participant. This tends to be more of an issue in studies related to questionnaires that measure characteristics of the participant, rather than objective stimuli. More relevant to this study is the issue of introspection and memory. Surveys are often distributed after a task is completed, and its accuracy is dependent on the ability of the participant to remember their experience during the study. Multiple research studies have shown that human memory is far from static. This can be dangerous if a researcher chooses to solely rely on self-report methods to test a hypothesis. We believe these self-report methods in tandem with biometric methods can help ensure the validity of the questionnaires, and provide information beyond the scope of self-report scales.

Research Questions
We know from previous research that the quality of websites mediates many aspects of e-

commerce, and provides insight as to how consumers view the webpages in general. However, simply knowing a webpage is perceived as lower quality doesn't give insight as to what aspects of a page are disliked by a user. Additionally, it's possible that the user is misremembering aspects of the webpage or being dishonest in their assessment. Using eye tracking metrics, galvanic skin response, and facial expression measures in tandem with a scale aimed at measuring visual design quality has a couple of identifiable benefits. Using both can potentially identify patterns amongst the biometric measures and the questionnaire, which would strengthen the validity of the results. More so, the eye tracking data has the potential to identify patterns amongst websites of lower or higher quality. If found, these patterns can be used to evaluate particular aspects of a page that are impacting the quality of a webpage. Overall, we are interested in answering two questions:

RQ1: Can attitudinal changes regarding substantial website redesigns be captured using biometric measures?

RQ2: How do biometric measures correlate with self-reported measures of visual appeal?

Answering these questions has the potential to provide a method of justification for design changes, ranging from minor tweak to complete rebrands. There is not an easy way for companies to quantitatively analyze visual design decisions. A method for doing so would help companies evaluate visual designs before implementation in order to cost-justify them. To this end, we hope to demonstrate that biometric measurements can be used with questionnaires to verify and validate potential design changes a company or organization might want to implement.

Methodology

We conducted a within-subjects study examining how the perceived visual appeal of websites is correlated with biometric measures. Participants reviewed multiple websites with varying levels of visual appeal while biometric responses were collected via eye-tracking, facial expression, and galvanic skin response equipment. Participants also provided qualitative data regarding the perceived visual appeal of the websites using a scale identified from relevant literature. To thoroughly select stimuli for our research, we implemented a two phase experimental design.

Phase 1

Our first phase of research involved creating a Qualtrics survey to evaluate stimuli for our eye tracking study in phase two. We identified 21 websites and gathered a static image of the site's current design, as well as a static image of one the site's previous designs. We found the previous site designs using "Wayback Machine" (archive.org/web). Therefore, we had two images for each website (a "before" and "after" versions), and a total of 42 stimuli to be evaluated with our survey. We were careful to choose websites we believed were not well known to avoid any confounding influence of familiarity on the users' visual design scores. Evaluating 42 individual stimuli take a long time for each participant to evaluate, so we divided the stimuli pairs into two groups. Participants who volunteered to complete the survey were randomly assigned to a group, and evaluated a total of either 20 images (10 sites) or 22 images (ll sites).

For the survey itself, we decided to present the survey stimuli using a method similar to our eye tracking study design. Each participant was assigned to one of the two stimuli groups, and then shown each webpage for only two seconds. After viewing the page they were prompted to rate the visual appeal of the websites using the Visual Aesthetics of Websites Inventory: Short version (VisAWI-S; Moshagen & Thielsch, 2012). The VisAWI-S is geared to assess four visual design qualities: simplicity, diversity, colorfulness, and craftsmanship. Each quality is represented as a question in the VisAWI-S scale, which can be viewed in Table 1. We also included a question to gauge participants' familiarity with the webpage, which can be viewed in Table 1 as well. We included this question to account for any unintended effects prior interaction with a site might have on visual design assessment. For all items, participants were asked about their agreement with the

statement using a 7 point Likert scale, with 1 being "strongly disagree" and 7 being "strongly agree".

For all participants, the first webpage presented was a "dummy" site, which was intended to familiarize the subject with the presentation of stimuli and the VisAWI-S questions. Following the dummy website, the remaining stimuli were presented in a randomized order to avoid unintended ordering effects. We selected the 10 websites with the largest differences in average VisAWI-S scores between the "before" and "after" versions to be used in phase two of our research.

Phase II

For phase two we designed our study to compare participants' self-reported visual design assessment of website images to their biometric responses to the images. As previously mentioned, the design of our phase two study and phase one initial survey were similar in order to obtain consistent data about visual design quality.

We recruited a total of 15 participants, using a convenience sample of classmates and friends. Our participants were comprised of 10 females and five males. We ensured all participants had adequate vision (natural or corrected) and did not have anything around their face that should prevent the facial expression software from being able to capture facial expressions.

At the beginning of each session, the participants were briefed on the nature of the experiment, the purpose, and the overall procedure of the study. We told participants that the purpose was to gather feedback on websites. We did not tell them that we were focusing on visual design so that their ratings on the VisAWI-S would not be influenced since we wanted their honest first impression of the websites. Similarly, while we informed participants of the data collection methods, we did not initially say that the webcam recording was being used to analyze their facial expressions as we did not want their knowledge of that to influence their natural expressions. Once the premise and details of the study had been communicated clearly and the participants confirmed their desire to continue, we began testing by calibrating the equipment to the participants.

To assess visual design, we presented the stimuli in a particular manner to ensure accurate data collection. Each website stimuli consisted of a block of four parts: a static image, a web page image, a grey screen, and a Qualtrics survey. First, participants were shown an image that mimicked television static for six seconds. This image was a scrambled version of the following web page image, and was intended to normalize participants' pupil dilation to the brightness or luminosity of the web page they were going to view next. Following the static screen, they were shown the image of the web page for two seconds. We limited the display time to two seconds so they could get a first impression focused solely on visual design, without having time to inspect further elements of the page. Next, they were shown a blank, grey screen for 6 seconds. Because galvanic skin response (GSR) has a slower response time,, we included the grey screen to ensure we captured all relevant GSR data before they moved on to the survey. Lastly, participants were taken to a survey where they were presented with the VisAWI-S scale questions to evaluate the website they were shown. Similar to the survey, participants were initially shown a "dummy" website to acclimate them to the experiment methodology, and then we presented the remaining blocks of stimuli in a randomized order.

Once the participant had completed the survey, we went through a debriefing process and fielded any questions the participant might had. We also elaborated on the data collection methods and informed them that the webcam recording we previously mentioned would be collecting facial expression data to be analyzed by the software.

Tools

For data collection in phase one, we used Qualtrics for survey creation and data collection. Social media sites, University of Texas mailing lists, and personal outreach were used to recruit survey participants. The eye

tracking data was gathered using the Tobii TX300 eye tracker. Additionally, a webcam in conjunction with the iMotions Afdex module was used to gather data on participants emotions and facial cues. Lastly, Shimmer3 GSR+ was used to gather information on participants skin conductance.

Data Analysis

Phase I

The purpose of phase one was to purposefully select our stimuli based on how our original 21 sites were rated by a broader audience. As mentioned before, because we had 21 sites, and each site had a before screen and an after screen, resulting in 42 sites on which we needed to gather data. We believed 42 individual screens were too much for survey respondents to rate in one sitting (without having effects of participants' fatigue), so we opted to split the stimuli into two near equal halves. Qualtrics randomly assigned each respondent to rate half of our stimuli. We obtained 30 participant ratings on the visual design of the first half of our stimuli, and 26 participants reviewed and rated the second set of stimuli. From these 21 stimuli pairs, we aimed to select the 10 websites with the most significant difference in visual design ratings between the before and after versions, which was determined using a few methods.

Familiarity and Vis-AWI-S Score

When initially selecting the 21 web pages from which to create our stimuli, we made an effort to choose sites we felt were not familiar to our participants. This way a participant's familiarity or experience with a website would be less likely to impact their visual design score. To determine if there was any relationship between familiarity and a web page's visual design score, we conducted a Spearman rank-order analysis on the phase one survey results. Our analysis yielded a statistically insignificant p-value ($r_s(147) = 0.12, p > .1$), indicating there was no significant relationship between the participants' familiarity and visual design scores of the stimuli.

Before/After Analysis

We examined how the ratings of visual appeal across all participants differed for each stimuli pair. Before we performed any statistical testing, we simply wanted a high-level look at how the before and after versions differed according to our survey. We first looked at the survey responses to calculate a participant's visual design rating for the stimuli by averaging their responses on the four VisAWI-S scale questions. Once completed for all participants, we averaged all participant responses for the before and after images for each website From there, we simply subtracted the two averages to find the difference between the before and after versions of the stimuli. We completed this analyses for all 21 stimuli pairs, noting the magnitude of each difference as well as the direction of the difference. The visual appeal difference of each stimuli can be seen in Table 1.

As a precautionary measure, we also conducted paired t-tests to determine if participants rated a before and after stimuli pair differently, and if so, to find the magnitude of the difference. We chose a paired t-test rather than an independent sample t-test because of our survey methodology. Although users did not see all 41 stimuli, they still viewed the before and after webpages of the stimuli they did examine. We made sure to account for this by using a paired sample t-test when analyzing the difference between web pages. Similar to our first analysis, a participant's rating of visual appeal was created by averaging their responses on the four VisAWI-s scale questions. Across the 21 stimuli, we noted which stimuli pair yielded the largest difference between the before and after pairs and noted the direction of the effect as well. All pairs can be viewed in Table 1, with their respective t-statistic and p-value.

The overall goal of the phase was to identify the 10 stimuli with the largest difference in visual design ratings based on phase one and use them in phase two. However, there was a slight difference in the 10 stimuli identified by the difference in average visual design compared to the 10 identified using paired sample t-tests. Specifically, the discrepancy existed between the travel blog web page, which was selected by

the t-test, and the hospital web page, which was selected based on difference in average score. To determine which web page we used in the next phase, we created box plots to look for two specific elements: variance and outliers. Variance signifies how much participants visual design scores varied for a web page, and outliers would reveal abnormal data points. Using these boxplots, which can be viewed in Figure 1, we concluded the hospital web page was a more appropriate stimuli pair to include in phase two because of its low variance and lack of outliers.

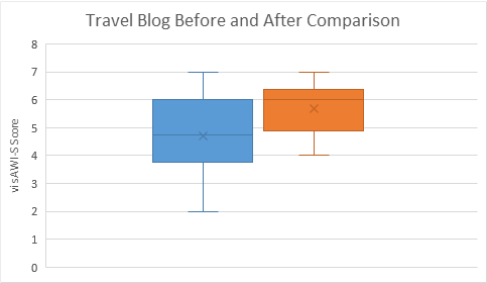

Figure 1 - Boxplot of travel blog web page scores

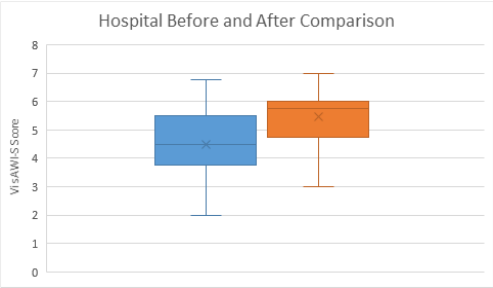

Figure 2 - Boxplot of hospital web page scores

Phase II

Our phase two analysis used a variety of non-parametric statistical tests. We selected tests and conducted data analysis with two goals in mind. First, we wanted to examine the relationship between our independent variable, the before and after pairs, and our dependent variables, visual design score, galvanic skin response peaks occurrence, pupil dilation, number of fixations, and average fixation duration. Additionally, we wanted to examine the overall relationship between a participants visual design score with our dependent variables. Unfortunately, we did not accurately capture facial expression data for all participants, and the data for the three participants we did collect did not produce enough results to analyze. Therefore, we have excluded the facial expression data from our analysis.

Dependent Variables

A number of actions had to be conducted to clean the data and prepare each variable for analysis. Visual design scores were calculated in the same manner as phase one: each respondent's answers on the four VisAWI-S questions were averaged. Galvanic skin response scores were measured in a unique way. For each stimuli, each participants had a binary data point which signified if they experienced a peak as they looked at a stimuli, with 1 being yes, and 0 being no. We looked at the frequency of how many of our 15 participants experienced a peak from a particular stimuli. The frequency of peaks per participants were incredibly varied and sporadic throughout our dataset, so using the frequency of a binary variable helped normalize the data. Pupil dilation looks at the average percent change from a participant's baseline pupil size to their pupil size while viewing a web page. A participant's baseline pupil size was determined by averaging their pupil size while they viewed a web page's respective static image. Additionally, looking at percent change, rather than raw pupil size change, allows us to compare each participant's pupil dilation with the other participants. The number of fixations is simply how many fixations occur as a participant views a stimuli. Lastly, average fixation duration, measured in milliseconds, looks at the the average length of time a participant fixates on a stimuli. A brief summary of each dependent variable can be found in Table 2.

Before & After Pairs

Despite the variety of variables, the methods used to examine the difference between the before and after stimuli pairs were largely the same. Although the order of our stimuli was randomized to avoid unintended ordering effects, every participant viewed both the before webpage and after page, which points to the need for a paired statistical test. Additionally, because our dependent variables violate the assumption of normal distribution, we determined Wilcoxon signed-rank tests were

appropriate to examine the dependent variables across the before and after webpages.

Across our comparisons of before and after webpage and our dependent variables we discovered a few comparisons that yielded statistically significant figures. As a validator of our phase one results, our Wilcoxon signed-rank test shows participants in phase two consistently rated the visual design of the before web pages (M = 3.38 , SD = 1.42) much lower than visual design of the after web pages (M = 5.40, SD = 0.97) , Z = -6.66, $p < 0.01$. Additionally, mean fixation duration showed a similar result. Average fixation duration amongst participants was significantly higher when participants viewed the before web pages (M =188.45, SD = 36.52) compared to the after web pages (M = 174.89 , SD = 25.73), Z = -2.78, $p < 0.01$. The final Wilcoxon analysis that yielded a significant result was the the comparison of pupil dilation across the before and after webpage images. The average pupil dilation for participants viewing the before images (M = 0.03, SD = 0.04) was smaller than those for the after web pages (M = 0.05 , SD = 0.04), Z = -3.79 , $p < 0.01$. All Wilcoxon signed-rank test Z-statistics and p-values can be viewed in Table 3.

VisAWI-S Score and Dependent Variables

Beyond comparing our dependent variables across before and after web pages, we wanted to examine the relationship between our dependent variables and participants' ratings of visual design. We chose to use Spearman's rank correlation to assess the relationship between our dependent variables and visual design scores because we are not analyzing normally distributed data. Overall, none of our dependent variables yielded statistically significant Spearman coefficients at our $\alpha = .05$ threshold, but some correlations were marginally significant and warrant further examination. Specifically, our Spearman rank-order analysis on the relationship between pupil dilation and visual design score ($r_s(147) = -0.1$, $p = 0.24$) and average fixation duration and visual design score ($r_s(147) = -0.13$, $p = 0.11$), both produced correlation coefficients close to .1, and p-values below .25. All Spearman rank-order analysis results and statistics can be viewed above in Table 3.

Results & Discussion

As mentioned, our analysis focused on two goals. First, we wanted to examine the difference, if any, in our dependent variables across the before and after pairs of our stimuli. Additionally, we examined the relationship between respondents' visual design scores of the stimuli and our dependent variables. Overall, we hope to gain some insight on how certain biometric measures may correlate with one's perception of visual design.

Wilcoxon Signed-Rank Analysis Results

For our before and after comparisons, we used Wilcoxon signed-rank tests because our study design utilized paired samples, and our dependent variables do not satisfy the assumption of normal distribution. First, we confirmed the findings of our phase one design by comparing the visual design scores from the before group to the after group, and found the before group was consistently rated lower (Z = -6.66, $p < 0.01$). From there, we found that participant's pupil dilation and average fixation duration significantly differed when they viewed a before web page versus an after web page. Contrarily, fixation frequency and GSR peak frequency yielded statistically insignificant Z-statistics (Table 1).

In regard to participants' pupil dilation, our significant result could mean a few things. On average, analysis showed our average participants pupil size changed by .02% when they viewed a before stimuli. This shows participants pupil size, on average, increased by 2% from a static screen to a before web page image. Conversely, average participants pupil size increased .05% from a static screen to an after web page image. In uncontrolled circumstances, an increase in pupil size could signify an adjustment to the brightness or luminosity of either a screen. However, the use of a static screen created from each web page image ensured our analysis accounted for this. This difference in pupil dilation could be explained by an increased level of engagement when participants viewed after web pages.

Research examining pupil size and task engagement revealed a relationship between the two variables (Broadway, Franklin, Mrazek, Schooler, & Smallwood, 2009). It is possible users found the after web pages to be visually engaging, as their layout and design is similar to current websites. On the other hand, the design of our before web page stimuli were archaic compared to many frequently visited web pages on the internet. As a result, users possibly were not as engaged when viewing the before stimuli. Facial recognition analysis could be useful in this scenario to observe participant engagement and emotions while viewing a web page. Having this information could provide additional context for data analysis, and provide a more complete picture of the relationship between biometrics and visual design assessment.

The significant difference between average fixation duration on before and after web pages was another effect that could be explained by previous research. Research looking at fixation duration and reading found a positive correlation with average fixation duration and passage difficulty (Ashby, Chace, Kathryn, & Slattery, 2006). Additionally, research has shown a positive correlation between participant memory load and fixation duration (Leeuwen, Meghanathan, & Nikolaev, 2015). The visual design of the before stimuli as compared to the design of commonplace websites on the web today (represented by the after images) might instill a sense of unfamiliarity or confusion when viewing the before images. In order to understand the unfamiliar layout or design of the before screen, participants may use more cognitive resources (Kahneman, 1973) compared to viewing the after screen, which relates to fixation duration. Using facial expression analysis to identify any emotions participants exhibit while viewing the before and after web pages could provide insight on their experience and emotions while viewing the stimuli. Also, with a longer exposure to the stimuli, we could retrospectively gather feedback from participants about their impressions of the stimuli and gain insights on visual design elements that may consistently be appealing or appalling across stimuli. However, a longer exposure of the stimuli may introduce other confounding variables, factors not related to visual design, that would need to be controlled for.

Correlational Analysis Results

We measured the correlative relationship between participant's scores on the VisAWI-S scale and our dependent variables using Spearman's rank correlation coefficient. As seen in Table 3, there were no significant correlations between our dependent variables of interest and visual design scores. However, the relationship between average fixation duration and visual appeal score produced a marginally significant coefficient value ($r_s(147) = -0.13$, $p = 0.11$). The relationship between visual design score and pupil dilation produced the second largest correlation coefficient ($r_s(147) = -0.1$, $p = 0.24$), but was still statistically insignificant at $\alpha = .05$. The lack of significant results was a bit surprising, especially given that the before and after comparisons yielded significant results. One explanation for this finding is the overall concepts the two tests measure. Wilcoxon's signed-rank test is used to uncover if two group's mean rank significantly differ from each other. However, Spearman's rho is used to assess the relationship between two variables. Even though both concepts are related, they are fundamentally different, and naturally may yield different results. More so, although the correlational analysis was not statistically significant, our data showed a positive relationship between both average fixation duration, pupil dilation and visual design score. It is possible with an increased sample size, our correlational strength and effect size would increase.

In the context of our mission, it seems as if pupil dilation and average fixation duration provide a path to using biometric measures as a complementary resource to visual design scales. However, our results are far from conclusive. The lack of correlational evidence between pupil dilation and average fixation duration with visual design score indicate the variables may not be as closely related as we hoped, and therefore should not be used to complement visual design assessment. However, based on previous research looking at reading

comprehension, memory load, and eye tracking (Broadway et al., 2009; Ashby et al., 2006; Balazs et al., 2017), more in depth analyses can be done to explore pupil dilation and average fixation duration. For one, retrospective methods can be used to gather qualitative data about what participants thought about while viewing a web page. Through this method, participants could give insight about why they fixated on certain areas of a page, or what when through their mind during moments where their pupils dilate. Additionally, analyzing and coding eye tracking videos can be used to uncover fixation patterns, and potentially explain visual design assessment.

Conclusion

By examining data from test subjects during a brief exposure to several websites, we hoped to explore the relationship between the self-reported evaluation of visual design quality and key biometric measurements of a subject's emotional valence and arousal. Subjects were exposed to ten pairs of websites before and after a substantial visual design change and asked to evaluate the website based on their initial impressions of the site's visual design quality using the VisAWI-S scale. During this assessment we collected GSR, facial expressions (limited by errors in initial study configuration), pupillary response, and fixation data using iMotions software coupled with a Tobii eye tracker, Shimmer GSR device, and Affdex facial expression analysis toolkit. This data was analyzed to discover relationships between the independent and dependent variables, as well as relationships between certain dependent variables.

Significant Findings

Upon data analysis, we discovered there is some evidence supporting the claim that biometric data can be used to validate the insights gained from the VisAWI-S self-report questionnaire. The most significant relationships were found in the eye-tracking fixation data. The number of fixations were fewer and total fixation duration was longer in before sites than after sites.
The average fixation duration and total number of fixations per stimulus were correlated inversely with one another, as would be expected due to the fact that more fixations results in less time spent fixated on any part of the image.

Lessons Learned

There are several key insights to be gained from this research. These insights can be categorized into certain insights arising from the data and analysis, and insights which concern the process in which this research was conceived, designed, and performed. The first class of insights are broad in scope and discussed more fully in the previous section. This second class of insights consists of broad lessons about the research process, as well as specific lessons about eye-tracking research on the iMotions platform.
Reflecting on the steps in which we undertook this study, it is clear that a significant amount of thought must be given to the nature of the data that will be collected and analyzed. As a result of choosing the specific collection of variables in this paper, data analysis required several disparate statistical analyses be performed. Each analysis provided an opportunity to learn about the ways in which different levels of measurement (ordinal, rank, interval) must be tested against one another. This was one of the most valuable aspects of conducting this particular study, as understanding these tools will be vital to understanding future research.
Additionally this study illuminated several features of the iMotions platform. The software has a moderately steep learning curve, requiring hours of configuration to arrive at a testable study. Webcam recordings must be enabled for all stimuli in order to obtain accurate Affdex data analysis post hoc. The limitations of the iMotions data analysis toolkit compounded the difficulty of our data analysis. Analyzing the raw data in third-party applications was initially difficult due to the size and complexity of the raw sensor data.
After conducting this research, we are more adequately prepared to use the iMotions platform in future research, as well as more comfortable with the research process in general.

Limitations

It is clear that this study is inadequate to fully understand the complicated relationship between a test subject's subjective rating of their first impressions of a website's visual appeal and the

biometric data obtained during the viewing of the website. However, it is clear that the relationship does exist. This study's failure to adequately capture facial expression data for the full cohort of test subjects is one obvious limitation of the research's findings. Another limitation is related to the fact that the study was conducted in an eye-tracking laboratory, with somewhat limited ecological validity.

When coupling the iMotions platform with the Shimmer GSR measurement device, Tobii eye tracker, and the Affdex facial expression analysis toolkit, we discovered some of the inherent limitations of these measuring devices and the measures themselves. Since GSR is a measure which has a large delay between stimulus and response, we attempted to design the study to account for this delay. In order to accurately measure pupillary response, we learned that pupil size must first be normalized to account for stimulus luminosity. The limited amount of data analyzed with the Affdex toolkit precludes a thorough understanding of its capabilities. However, it seems likely that the particular design of this study was not suited to facial expression analysis due to the fact that there was only one recorded frame in which Affdex registered an emotion. Failure to record emotions might be due to an Affdex configuration error, or because of the lack of emotionally stimulating content in the study.

Future Directions & Next Steps

Continuing onward, future research would likely seek to more fully explore the relationship between biometric measurements and self-reported measures of visual design quality like the VisAWI-S. New studies which analyze a larger pool of facial expression data would be able to support conclusions which this study alone could not, due to the lack of facial expression data collected. Additionally, the VisAWI-S instrument is only VisAWI-S tool is only one of many used to assess visual design and aesthetic in the literature. Having participants analyze web pages using other scales may show a stronger relationship with physiological responses.

The applications of this study's methodology and analysis techniques can be extended outside the domain of visual design. Additional aspects of a website's user experience such as information architecture, content strategy, or interaction design could be examined using similar techniques as presented here to gain further insights into the validity of traditional subjective measurements used to study these other aspects of user experience.

Works Cited

Rayner, Keith, et al. "Eye movements as reflections of comprehension processes in reading." *Scientific studies of reading* 10.3 (2006): 241-255.

Jones, C., & Kim, S. (2010). Influences of retail brand trust, off-line patronage, clothing involvement and website quality on online apparel shopping intention: Online apparel shopping intention. *International Journal of Consumer Studies*, *34*(6), 627–637. https://doi.org/10.1111/j.1470-6431.2010.00871.x

Kim, S., & Stoel, L. (2004a). Apparel retailers: website quality dimensions and satisfaction. *Journal of Retailing and Consumer Services*, *11*(2), 109–117. https://doi.org/10.1016/S0969-6989(03)00010-9

Kim, S., & Stoel, L. (2004b). Dimensional hierarchy of retail website quality. *Information & Management*, *41*(5), 619–633. https://doi.org/10.1016/j.im.2003.07.002

Kahneman, Daniel. *Attention and effort*. Vol. 1063. Englewood Cliffs, NJ: Prentice-Hall, 1973.

Kuan, H.-H., Bock, G.-W., & Vathanophas, V. (2008). Comparing the effects of website quality on customer initial purchase and continued purchase at e-commerce websites. *Behaviour & Information Technology*, *27*(1), 3–16. https://doi.org/10.1080/01449290600801959

Lin, H.-F. (2007). The Impact of Website Quality Dimensions on Customer Satisfaction in the B2C E-commerce Context. *Total Quality Management & Business Excellence*, *18*(4), 363–378. https://doi.org/10.1080/14783360701231302

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. https://doi.org/10.1080/01449290500330448

Meghanathan, Radha Nila, Cees van Leeuwen, and Andrey R. Nikolaev. "Fixation duration surpasses pupil size as a measure of memory load in free viewing." *Frontiers in human neuroscience* 8 (2015): 1063.

Muzellec, L., & Lambkin, M. (2008). Corporate Rebranding and the Implications for Brand Architecture Management: The Case of Guinness (Diageo) Ireland. *Journal of Strategic Marketing*, *16*(4), 283–299. https://doi.org/10.1080/09652540802264124

Roy, S., & Sarkar, S. (2015). To brand or to rebrand: Investigating the effects of rebranding on brand equity and consumer attitudes. *Journal of Brand Management*, *22*(4), 340–360. https://doi.org/10.1057/bm.2015.21

Stojmenovic, M., Pilgrim, C., & Lindgaard, G. (2014). Perceived and objective usability and visual appeal in a website domain with a less developed mental model (pp. 316–323). ACM Press. https://doi.org/10.1145/2686612.2686660

Wells, Valacich, & Hess. (2011). What Signal Are You Sending? How Website Quality Influences Perceptions of Product Quality and Purchase Intentions. *MIS Quarterly*, *35*(2), 373. https://doi.org/10.2307/23044048

Appendix

Table 1. Items included in the Vis-AWI-S instrument

Factor	Item
Simplicity	Everything goes together on the site.
Diversity	The layout is pleasantly varied.
Colorfulness	The color composition is attractive
Craftsmanship	The layout appears professionally designed
Familiarity*	I am familiar with this website

** question is simply to gauge familiarity for the study, and is not part of the Vis-AWI-S instrument*

Note: Participants were asked about agremeent with the item using a 7-point likert scale

Table 2. Descriptive Statistics, Mean Difference, and p-values for Website Stimuli

	Before		After			
Website	*M*	*SD*	*M*	*SD*	Mean Difference	*p*
Joy Kitchen	3.49	1.30	5.61	0.93	2.12	0.00
Seacom	3.27	1.59	5.35	1.20	2.08	0.00
Food Blog	3.59	1.30	5.59	0.80	2.00	0.00
Credit Union	3.29	1.26	5.18	1.07	1.89	0.00
Travelers	3.61	1.39	5.38	1.24	1.78	0.00
Sporcle	4.23	1.23	2.45	1.12	-1.78	0.00
Eagle	3.93	1.47	5.45	0.82	1.52	0.00
Oberlin	4.00	1.25	5.47	0.84	1.47	0.00
Valve	3.88	1.56	5.10	1.42	1.22	0.00
Hospital*	4.47	1.33	5.48	0.85	1.01	0.00
Travel Blog*	4.71	1.23	5.69	1.01	0.98	0.00
Space	4.35	1.55	5.29	1.09	0.94	0.00
School	5.04	1.44	5.63	0.80	0.60	0.06
Book Publisher	5.12	1.27	5.63	1.17	0.51	0.10
Sneakers	4.78	1.37	5.20	1.34	0.42	0.14
Stance	5.08	0.88	5.41	0.95	0.33	0.09
City	4.79	1.18	5.12	0.88	0.32	0.07
IEEE	3.95	1.30	4.26	1.40	0.31	0.24
Rise	5.08	1.00	4.89	1.27	-0.18	0.30
Audio Technica	3.94	1.52	4.05	1.37	0.11	0.71
Bloomberg	3.63	1.35	3.52	1.26	-0.11	0.73

Note: Stimuli are ranked by largest to smallest absloute mean difference.

** indicates the stimuli that differed between our paired sample t-test and mean difference comparisons.*

Table 3. Description of Dependent Variable

Variable	Description
Visual Design Score	A measure of visual design, calculated by averaging a participants responses on the four VisAWI-S questions
Pupil Dilation	A measure of a participants average pupil dilation from the baseline to the stimuli.
GSR Frequency	A measure of how many participants per stimuli experienced a peak
Fixation Frequency	A measure of how often a participants fixates on a area of the stimuli
Average Fixation Duration	A mean of how long a participant fixates on on areas of the stimuli. Measured in milliseconds (ms)

Table 4. Wilcoxon Signed-Rank and Spearman's Rank Correlation Analysis

	Change in Pupil Size	GSR Peaks	Fixation Frequency	Mean Fixation Duration	Visual Design Score
Visual Design Score	$r_s = -0.1$	$r_s = 0.039$	$r_s = 0.02$	$r_s = -0.13$	
Before/After	$Z = -3.79$*	$\chi^2 = 0.97$	$Z = -0.56$	$Z = -2.78$*	$Z = -6.66$*

* $p < .01$