INTRODUCTION OF A NEW PSYCHOMETRIC TOOL TO MEASURE USER RESPONSES IN A KELLY BLUE BOOK DESIRABILITY STUDY

A Graduate Project submitted in partial fulfillment of the requirements
For the degree of Master of Arts in Psychology,
Human Factors and Applied Psychology

By
JonDelina Buckley

May, 2012
The Graduate Project of JonDelina Buckley is approved:

________________________________
Dr. Jill Strawbridge, Ph.D.

________________________________
Barry Berson

________________________________
Dr. Tyler Blake, Ph.D., Chair

California State University, Northridge
# TABLE OF CONTENTS

SIGNATURE PAGE ...........................................................................................................II
ABSTRACT............................................................................................................................IV
INTRODUCTION ..................................................................................................................1
BACKGROUND OF USING QUESTIONNAIRES FOR USABILITY STUDIES ........ 4
COMPARING QUESTIONNAIRES FOR ASSESSING WEBSITE USABILITY .......... 7
THE DEVELOPMENT OF THE SYSTEM USABILITY SCALE......................... 9
DEVELOPMENT OF THE SUPR-Q........................................................................... 14
COMPARISON OF THE SUS AND THE SUPR-Q ............................................. 19
USING THE SUPR-Q FOR THE KELLEY BLUE BOOK DESIRABILITY STUDY.. 22
  METHOD ......................................................................................................................... 22
  PARTICIPANTS ............................................................................................................... 28
  ANALYSIS ...................................................................................................................... 30
  DISCUSSION ................................................................................................................... 37
REFERENCES .................................................................................................................. 46
ABSTRACT

INTRODUCTION OF A NEW PSYCHOMETRIC TOOL TO MEASURE USER RESPONSES IN A KELLY BLUE BOOK DESIRABILITY STUDY

By

JonDelina Buckley

Master of Arts in Psychology,
Human Factors and Applied Psychology

Psychometrics is the field of study concerned with the theory and technique of educational and psychological measurement. This includes the measurement of knowledge, abilities, attitudes and personality traits. Psychometric measurement tools are often used in the User Experience profession to assess user’s perceptions of their interactions with websites, software and hand held devices. Various psychometric questionnaires have been employed for assessing the usability of interactive systems, (e.g., Questionnaire for User Interface Satisfaction (QUIS), Computer System Usability Questionnaire (CSUQ), etc.) However, one of the most popular psychometric measurement tools for Human Factors engineers and User Experience professionals is the System Usability Scale (SUS). The SUS was developed in 1986 by John Brooke as a “quick and dirty” usability scale during his tenure at the software company Digital Equipment Corporation (DEC).

It has (arguably) been established that the SUS is a highly effective tool for use when measuring usability and “user friendliness”. However, when User Experience and Human Factors professionals need to measure more emotional reactions to designs such as
perceived trust, comfort, loyalty and appearance what psychometric tool alternatives exist? Recently, a new psychometric questionnaire has been introduced in the field.

Labeled the Standard Usability Percentile Rank Questionnaire (i.e., SUPR-Q), this new psychometric tool’s global score is made up of a percentile ranking based on a composite weighting of four factors: Usability, Trust & Credibility, Appearance and Loyalty. In addition, percentile ranks are provided for each factor so the strengths and weakness by factor can be assessed (Sauro, 2010).

In November of 2010, the User Experience Research team for Kelley Blue Book, a third party automotive website, used the SUPR-Q as the post-test questionnaire for a Desirability Study. The SUPR-Q was used to assess measures of Trust, Comfort, Loyalty and Appearance of two proposed site redesign concepts. Ultimately, the KBB.com User Experience design and research teams discovered that the new SUPR-Q psychometric questionnaire results did provided valuable insights. Particularly enlightening were percentile rank comparisons across all four sub-categories (Trust, Comfort, Loyalty and Appearance) against other competitive third party auto sites. However, other confounding factors may have rendered the SUPR-Q less effective in providing statistically significant data regarding how the two proposed redesign concepts performed against each other.
Introduction

In November of 2010, the User Experience Research team for Kelley Blue Book conducted the company’s first Desirability study. Kelley Blue Book is an eighty-five year old third party automotive company that transitioned from a printed soft cover edition of its iconic blue hued used car trade-in “valuation” guide to an online website in 1995.

In early 2010, a long-awaited redesign of the company’s website, KBB.com had finally been approved. The KBB.com website had not undergone a major redesign in nearly five years. The creative team was interested in assessing whether a more modern look and feel would resonate with the sites predominately Baby Boomer audience while simultaneously appealing to a younger audience the company hoped to acquire. To achieve this goal, a Desirability study research method was decided upon. First introduced in 2002 by Trish Miner and Joey Benedek of Microsoft, Desirability studies utilize “product reaction cards.” Product reaction cards leverage descriptive terms to aid users in describing their feelings, reactions and perceptions about the designs of software, handheld and website products. However, a Desirability study should not be confused with traditional market research related to gathering users’ perceptions of a brand. Desirability studies often explore users’ perceptions regarding how well proposed designs might support users potential interactions with a site. Desirability studies often leverage unbranded concepts and are generally conducted using an attribute selection methodology. This attribute selection methodology can inform design teams in several ways. It can be used to assess how users perceive different potential designs for a site by having them select which attributes they associate with each design. A Desirability study
can also provide insights as to why different design directions evoke certain responses in the target audience. Further, a Desirability study can precisely measure visual design directions against specific adjectives (such as brand attributes) to help support a final direction (Benedek, Miner, 2002).

To augment the standard Miner and Benedek Desirability study experiment design, the Kelley Blue Book User Experience Research team decided to use a recently developed psychometric tool, called, the SUPR-Q as the post-test questionnaire. The SUPR-Q was used to assess measures of Trust, Comfort, Loyalty and Appearance of two proposed site redesign concepts. This project explored three main questions related to that research. First, when compared to the more often used System Usability Scale (SUS), how did the SUPR-Q perform? Second, why did the Kelley Blue Book User Research team choose the SUPR-Q over the SUS? Finally, we attempted to answer the question, what benefits could using the SUPR-Q provide over the SUS?

To set the proper framework for an analysis of a new psychometric measurement tool, this project first provides a brief look back at the use of psychometric tools in the world of Human Factors and User Experience. Additionally, we discuss the reasons and rationale for using a quantitative measurement tool (like a psychometric questionnaire) and the need for precision in the experiment design, especially related to adequate sample sizes. We also investigate other research in the field comparing psychometric measurement tools. Next, we look at the process used to develop the most often used psychometric tool, the SUS, and how the development of the SUPR-Q compares to this process. Further, we assess progress in the development of psychometric tools used to measure emotion-based user feedback (as in a Desirability study) in relation to various products and services. Finally, we provide suggestions for the future development of
potential tools that attempt to assess the fuzzy boundaries of users’ emotions and feelings toward software, hand held and website products.
Background of Using Questionnaires for Usability Studies

In his paper discussing the use of questionnaire methods for usability studies, J. Kirakowski notes a reference from Anastasi in her classic book on the subject (Anastasi, 1968). Anastasi notes that a psychological questionnaire is essentially an objective and standardized measure of a sample of behavior (Kirakowski, 1994). Kirakowski summated that using a psychological questionnaire to evaluate a computer system by measuring the behavioral disposition of end users towards it was an “obvious step.”

Further, according to Kirakowski, in 1978, Dzida, Herda and Itzfeldt reported the first large-scale questionnaire specifically addressed to the problem of devising a rigorous measure of user-perceived quality. They postulated that user-perceived quality would turn out to be a multi-dimensional concept, but additionally made the further assumption that, each dimension of user-perceived quality is defined by a set of system properties each of which is mainly associated with this dimension (Dzida et al., 1978).

Dzida et al., produced a seven-factor structure derived from the ISO9 9241, Part 10 sample of one hundred system requirements for user-perceived quality. The researchers started from a consideration set of system characteristics (such as input format, response time, detail of explanation, etc.) as opposed to users’ expectations and attitudes toward the system being evaluated (Kirakowski, 1994). Many aspects of Dzida et al.’s list of software quality features are found in the EVADIS evaluation procedure. The EVADIS is a comprehensive evaluation approach that takes the tasks, the user and the organizational context into consideration during the evaluation process, and provides computer support for the use of the evaluation procedure (Opperman, et al., 1988). Such lists place considerable importance on the presence or absence of desirable software features which,
on the judgment of experts, are considered to contribute to user satisfaction (Kirakowski, 1994). In 1991, Lewis developed a three-item questionnaire called the After Scenario Questionnaire (ASQ). The ASQ was developed to be used immediately following scenario completion in scenario-based usability studies, where a scenario, according to the author, is a collection of related tasks (Lewis, 1991). The objective of the three questions of ASQ was to measure one single underlying aspect of users' perceptions of how easily and quickly the scenarios were completed and the contribution of support information to carrying out the tasks. In a 2009 study analyzing the tool, Jeff Sauro and Joe Dumas maintained:

The ASQ exhibited acceptable reliability and sensitivity and the Likert format was easy for participants to use and easy for researchers to score. A more recent study compared four variations of the Likert question type, including two of the ASQ questions. In that study, each participant used one of the formats to rate tasks. All of the scales had significant correlations with task time and a post-test SUS questionnaire (Dumas and Sauro, 2009).

Work on specific questionnaire methods of analyzing user reactions was started in the Human Factors Research Group (HFRG), University College Cork, in 1986. The focus of the research was to examine user reactions to a specific computer product with which the user sample had some experience. The first result from the HFRG studies was the Computer User Satisfaction Inventory (CUSI). CUSI is a short questionnaire of 22 items. Two subscales of usability were established, called, at the time, Affect (the degree to which users like the computer system) and Competence (the degree to which users feel supported by the computer system). These subscales were arrived at through cluster analysis of inter-correlation responses to individual questions in a large initial item pool. This item pool was gathered from reviewing the literature and discussions with end-users about their reactions when carrying out their normal tasks on their usual systems. The
range of systems sampled was large and heterogeneous (Kirakowski, 1987, Kirakowski and Corbett, 1988). Initial estimates of the reliability of the CUSI questionnaire on a sample of data independent of the sample on which the original dimensions had been discovered, showed an overall reliability of 0.94. Two separate scales showed reliabilities of 0.91 for Affect, and 0.88 for Competence. At about the same time that CUSI was published, Shneiderman included the QUIS (Questionnaire for User Interface Satisfaction) scale in his book “Designing The User Interface”, published in 1987. An important evaluation of QUIS was published by Chien, Diehl and Norman in 1988, by which time the questionnaire had been incremented up to version 5.0. The questionnaire was included in the second edition of Shneiderman's 1992 book, Designing the User Interface: Strategies for Effective Human-Computer Interaction, (Kirakowski, 1994).

QUIS version 5.0 included one introductory section, which provided a scale to measure users’ overall reaction to the software, and four other sections, each consisting of between 4 to 6 items. These sections were: “Screen,” “Terminology and System Information,” “Learning,” and “System Capabilities”. Chin, Diehl and Norman reported a high reliability coefficient of 0.94 for the entire scale. The items within the sections were fairly specific to each section's theme and therefore strong inter-correlations between items within the sections were expected (i.e., items such as “learning to operate the system,” “exploring new features by trial and error”, and “remembering names and use of commands” were found in the “Learning” section) (Kirakowski, 1994).
Comparing Questionnaires for Assessing Website Usability

Two studies carried out in 1990 and 1991 examined three questionnaires, CUSI, SUS and QUIS version 5.0. Wong and Rengger (1990) calculated correlation coefficients between CUSI Affect, CUSI Competence, SUS, and the overall QUIS score. They found that CUSI Affect, SUS, and QUIS all correlated together quite well with coefficients of between 0.672 to 0.744. CUSI Competence however correlated less strongly with these other scales: 0.584 with SUS and 0.379 with QUIS. The authors concluded that CUSI Affect, QUIS and SUS were, most probably, measuring a dimension which could be called Affect, whereas CUSI Competence, with its low correlations with all the other tests, may well be measuring something else (Kirakowski, 1994).

In 2004, Tullis and Stetson, from the Usability labs at Fidelity Investments, conducted a psychometric comparison tool study with 123 participants to answer the question, “How well do the various questionnaires apply to the assessment of websites?” The study compared five psychometric questionnaires, the System Usability Scale (SUS), the Questionnaire for User Interface Satisfaction (QUIS), the Computer System Usability Questionnaire (CSUQ), Microsoft’s Product Reaction Cards and a special psychometric questionnaire developed for use in the usability labs at Fidelity Investments. Participants used one of the five standard usability questionnaires to rate the usability of two websites. With the entire sample size, all five questionnaires indicated superior usability for the same website. Because at the time, no practical usability test would have had such a large number of participants, they conducted a Monte Carlo simulation (a technique that involves using random numbers and probability to solve problems). The Monte Carlo analysis was used to examine if, as the sample size increased from 6 to 14, which of the
questionnaires would converge most quickly to the “correct” conclusion regarding the differences between the websites’ usability. “Correct” was defined as a significant t-test consistent with the decision reached using the total sample size. The end result showed that two of the questionnaires, the SUS and the CSUQ, met this goal the most quickly. Both psychometric questionnaires made the correct decision over 90% of the time when n=12 (Stetson and Tullis, 2004).
The Development of the System Usability Scale

In the original paper published by Brooke in 1986 discussing the System Usability Scale (SUS), the author describes the tool as a “quick and dirty” usability scale (see Figure 1.1).

The SUS consists of 10 items, with odd-numbered items worded positively and even-numbered items worded negatively. The SUS is administered by presenting the items to participants as 5-point scales numbered from 1 (anchored with “Strongly disagree”) to 5 (anchored with “Strongly agree”). If a participant fails to respond to an item, it is assigned 3 (the center of the rating scale). After completion, each item’s score contribution is determined, ranging from 0 to 4. For positively-worded items (1, 3, 5, 7 and 9), the score contribution is the scale position minus 1. For negatively-worded items

<table>
<thead>
<tr>
<th>Item</th>
<th>Strongly Disagree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I think that I would like to use this website frequently.</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>2. I found this website unnecessarily complex</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>3. I thought this website was easy to use.</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>4. I think that I would need assistance to be able to use this website</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>5. I found the various functions in this website were well integrated.</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>6. I thought there was too much inconsistency in this website</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>7. I would imagine that most people would learn to use this website very quickly.</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>8. I found this website very cumbersome/awkward to use.</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>9. I felt very confident using this website.</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>10. I needed to learn a lot of things before I could get going with this website.</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>
(2, 4, 6, 8 and 10), the score contribution is 5 minus the scale position. To get the overall SUS score, the sum of the item score contribution is multiplied by 2.5. Thus, SUS scores range from 0 to 100 in 2.5-point increments.

Originally, limited information was known or published about the process for developing the SUS. In the ensuing twenty-five years, this self-described “quick and dirty” usability scale has become a popular questionnaire among Usability Professionals and Human Factors engineers for post-test subjective assessment of usability. Research conducted on the SUS has shown that although it is fairly quick, it is probably not all that dirty. The ten SUS questions were developed by selecting from a pool of 50 potential items, based on the responses of 20 people who used the full set of items to rate two software systems, one of which was relatively easy to use, and the other relatively difficult. The items selected for the SUS were those that provided the strongest discrimination between the systems (Lewis and Sauro 2009). The typical minimum reliability goal for questionnaires used in research and evaluation is .70. An early assessment of the reliability of the SUS based on 77 cases indicated a value of .85 for coefficient alpha (a measure of internal consistency often used to estimate reliability of multi-item scales) (Lewis and Sauro 2009). In 2008, Bangor, Kortum, and Miller, in a study of 2,324 cases, found the coefficient alpha of the SUS to be .91. Bangor et al. also provided some evidence of the validity of the SUS, both in the form of sensitivity (detecting significant differences among types of interfaces and as a function of changes made to a product) and concurrent validity (a significant correlation of .806 between the SUS and a single 7-point adjective rating question for an overall rating of “user friendliness”) (Bangor et al., 2008). In 2009, a paper published by Lewis and Sauro outlined the process for developing the System Usability Scale (SUS). It revealed that in
fact, the SUS was created using the standard methodology for developing most
psychometric questionnaires. The five step process consists of the following stages:
Identify the Constructs, Generate Candidate Questions, Pilot Test, Winnowing,
Reliability and Validity (Lewis and Sauro, 2009).

In the “Identify the Constructs” step, the researcher first identifies the constructs to
measure, (i.e., Usability, Ease of Use, Satisfaction and Ease of Learning). In the second
step, “Generate Candidate Questions,” the researcher then comes up with several
questions that attempt to measure the constructs of interest. The questions can be
completely made up, based on existing questions from other published research or some
combination of both. The main goal for this second step is to create a large set of
candidate questions. The SUS started with fifty items. In step three, “Pilot Test”, the
candidate questions are tested with the target audience. To appropriately analyze the
collected data, the ideal number of responses for each question ranges from 150 to 400.
However, a modest sample of less than 50 can effectively identify poor items. The more
constructs measured, the larger the sample size required. In the “Pilot Test” step, the goal
is to identify which questions do not effectively measure the constructs and to see which
questions tend to cluster together. The original 50 SUS candidate questions were used on
two systems by just 20 people. Step four consists of the “Winnowing” process. This
step’s goal is to determine which questions effectively measure the construct. The
process for this step involves excluding questions which have low correlations with the
total average and with each other. The multivariate statistical method, factor analysis,
reveals which questions cluster together. These clusters (or factors) become the sub-
scales of the questionnaire. Parenthetically, it may be valuable to note that the SUS was
not designed to have any subscales and so was not subjected to factor analysis during its
development (Lewis and Sauro, 2009). In the original paper by Brooke, he reported a strong correlations among the selected items (absolute values of $r$ ranging from .7 to .9), but did not report any measures of reliability or validity, referring to the SUS as a “quick and dirty” usability scale. For these reasons, he cautioned against assuming that the SUS was any more than a unidimensional measure of usability.

SUS yields a single number representing a composite measure of the overall usability of the system being studied. Note that scores for individual items are not meaningful on their own (Brooke, 1986).

Given that the SUS was originally developed with data from only 20 participants, this caution was appropriate. However, additional research conducted by Lewis and Sauro in 2009 revealed that the SUS actually has two factors that have reasonable reliability. The researchers conducted a factor analysis of two independent SUS data sets and discovered that the SUS actually has two factors, Usability (8 items) and Learnability (2 items). These new scales have reasonable reliability (coefficient alpha of .91 and .70, respectively). They correlate highly with the overall SUS ($r = .985$ and .784, respectively) and correlate significantly with one another ($r = .664$), but at a low enough level to use as separate scales. A sensitivity analysis using data from 19 tests had a significant Test by Scale interaction, providing additional evidence of the differential utility of the new scales (Lewis and Sauro, 2009).

The final step in the development of a psychometric questionnaire is the assessment of “Reliability and Validity” for the tool. Once there are a set of subscales, developers typically assess reliability (repeatability) using a measure of internal-reliability called coefficient alpha or Cronbach’s alpha. Cronbach’s alpha provides a measure of how consistently users have responded to items in the questionnaire. The highest possible
score is 1.00, with .70 considered to be the lower boundary of acceptable internal reliability. If the total questionnaire or subscales have low internal reliability (Cronbach’s alpha < .70) then questions and subscales are added and removed until the reliability becomes acceptable. Where reliability measures the repeatability of the questionnaire results, validity is the extent to which the questionnaire is measuring what it is intended to measure (in this case, perceived usability). Validity is typically measured via correlation (correlating scores of the questionnaire with other measures of usability including other known valid questionnaires). Additionally, sensitivity, (ability to discriminate well between usable and unusable systems) is another important factor of validity. SUS was shown to discriminate well between a known easy-to-use linguistic tool and an unusable programming tool (Lewis and Sauro, 2009).
Development of the SUPR-Q

The creator of the SUPR-Q, statistician and User Experience professional, Jeff Sauro, describes the SUPR-Q as a psychometric rating scale constructed with the intent of measuring perceptions of Usability, Trust, Credibility, Appearance and Loyalty for websites.

The SUPR score is made up of a global score which is a percentile ranking based on four factors: Usability, Credibility (which includes Trust, Value & Comfort), Appearance and Loyalty. Additionally, each subscale has a percentile rank providing the capability to see the strengths and weakness by subscale. The SUPR-Q questionnaire contains thirteen standardized questions which generate a normalized score. The scoring system is based on the relative ranking from a database of approximately 112 websites (Sauro, 2010).

The SUPR was developed through the standard process of psychometric validation. Development of the SUPR-Q included over 75 candidate questions, tested on thousands of users across hundreds of websites to arrive at the 13 questions which provide the most reliable, sensitive and valid picture of the attitudes which imply the success of a website, (Sauro, 2010).

The SUPR-Q psychometric tool contains the following questions and subscales:

Usability

This website is easy to use.

I am able to find what I need quickly on this website.

I enjoy using the website.

It is easy to navigate within the website.

Credibility (Trust, Value & Comfort)

I feel comfortable purchasing from this website.

This website keeps the promises it makes to me.
I can count on the information I get on this website.

I feel confident conducting business with this website.

The information on this website is valuable.

**Loyalty**

How likely are you to recommend this website to a friend or colleague?

I will likely visit this website in the future.

**Appearance**

I found the website to be attractive.

The website has a clean and simple presentation.

There are five response options for twelve of the questions and can be displayed in a grid like the example below (See Figure 1.2).

<table>
<thead>
<tr>
<th>Question</th>
<th>Strongly Disagree 1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Strongly Agree 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>This website is easy to use.</td>
<td>☐</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>It is easy to navigate within the website.</td>
<td>☐</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>I enjoy using the website.</td>
<td>☐</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>I feel comfortable purchasing from this website.</td>
<td>☐</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>I am able to find what I need quickly on this website.</td>
<td>☐</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>I can count on the information I get on this website.</td>
<td>☐</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>I found the website to be attractive.</td>
<td>☐</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>I feel confident conducting business with this website.</td>
<td>☐</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>The website has a clean and simple presentation.</td>
<td>☐</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>The information on this website is valuable.</td>
<td>☐</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>This website keeps the promises it makes to me.</td>
<td>☐</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>I will likely return to this website in the future.</td>
<td>☐</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
</tbody>
</table>

Figure 1.2 Questions 1-12 of the SUPR-Q
The final question “How likely are you to recommend this website to a friend or colleague?” (called the “Net Promoter Score”) has 11 response options and can be displayed like the example below (See Figure 1.3). It is constructed as a 0-10 point Likert-type scale. The scale design requires the end-points and neutral middle option to be labeled.

![Figure 1.3 Questions 1-12 of the SUPR-Q](image)

Further, in developing the SUPR-Q as a percentile rank rather than a simple percentage score like the SUS, Sauro notes:

The SUPR-Q score is expressed as a percentile (hence percentile rank), so a SUPR-Q score of 75% would mean the global score is higher than 75% of all websites tested in the database. The same percentile ranking procedure is then done for all of the four subscales and by each of the 13 questions. This can also be done within a specific industry. The tool is constructed to have a range of rankings for each of the four subscales, allowing the ability to reveal potential differences in subscale rankings (i.e., some websites might score high in usability but low with respect to loyalty) (Sauro, 2010).

To support the validity of the SUPR-Q’s measures of Usability, Credibility, Trust and Appearance, the tool was compared to other known instruments which measure similar constructs. These comparison’s revealed that the SUPR-Q global score and usability score had a very strong correlation with the System Usability Scale (SUS) $r > .96$, $p < .01$. Additionally, an experiment was conducted with 108 users across 8 different websites in which participants completed both the SUPR-Q and the WAMMI questionnaire. The usability factor and global SUPR-Q score had a strong correlations $r > .8$, $p < .01$ with the average responses from the 20 WAMMI questions. A combination of
just four questions from the SUPR-Q accounted for 80% of the variability in the 20 WAMMI questions. The SUPR-Q was able to discriminate well against websites which have been known to have poor Usability, Trust, Loyalty and Appearance against others which score highly on these attributes (Sauro, 2010).

Reliability was assessed for the SUPR-Q tool by gathering data on its use on over 200 websites and by over 4,500 users from 13 countries (90% US & Canada). As part of the validation of the instrument, internal reliability was measured by using Cronbach’s alpha, inter-items correlation, item total correlation, ANOVA and factor analysis. Validity for usability was measured using Cronbachs’s alpha, and Coefficient alpha. Discriminate validity, (i.e., was it able to discriminate against high vs. low usability) was determined by looking at websites that are known to have high usability (i.e., Amazon) versus low usability websites (i.e., government websites) (Sauro, 2010).

The global score for each sub-scale exhibited strong internal reliability (see Figure 1.4).

<table>
<thead>
<tr>
<th>Factor</th>
<th>Cronbach Alpha</th>
<th>Mean. Inter-Item Correlation</th>
<th>Min. Inter-Item Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usability</td>
<td>0.937</td>
<td>0.788</td>
<td>0.724</td>
</tr>
<tr>
<td>Trust, Value &amp; Comfort</td>
<td>0.887</td>
<td>0.614</td>
<td>0.463</td>
</tr>
<tr>
<td>Appearance</td>
<td>0.82</td>
<td>0.698</td>
<td>0.698</td>
</tr>
<tr>
<td>Loyalty</td>
<td>0.715</td>
<td>0.76</td>
<td>0.76</td>
</tr>
<tr>
<td>Overall</td>
<td>0.935</td>
<td>0.593</td>
<td>0.391</td>
</tr>
</tbody>
</table>

Figure 1.4 Reliability for the Global Scores of the SUPR-Q
In reference to questions related to weighting in the SUPR-Q construction, Sauro notes:

All SUPR-Q subscales are equally weighted (i.e., no intentionally weighting mechanism was built in) across all 5pt items, except question number thirteen ("The Net Promoter Score" question) (Sauro, 2010).

Special consideration was given to the addition of question number thirteen ("The Net Promoter Score") in the development of the SUPR-Q and the subsequent analysis of the tools results.

The contribution of the NPS to the total overall SUPR-Q score is cut in half (i.e., only half as many points for the NPS score) to avoid disproportional weighting against all the other subscales. Users of the SUPR-Q are able to score the responses for the first 12 questions and add this to 1/2 the score for the last questions (question number 13, the Likelihood to Recommend question). This score can then be compared to the industry benchmarks. The lowest possible score is a 12 and the maximum possible score is a 65 (Sauro, 2010).
Comparison of the SUS and the SUPR-Q

When comparing the SUPR-Q tool to the SUS it must be noted that while both tools can measure usability factors they have some inherent differences. The System Usability Scale (SUS) was constructed to measure users’ perceptions of usability. Additionally, its questions are generic enough to allow it to be used on a wide variety of software and hardware applications. However, as a result of this generic quality, some of the tool’s questions are not as suited to websites (e.g., needing the support of a technical person). Further, while the SUS provides a reliable picture of users’ attitudes toward usability, it is limited to only perceptions of usability and excludes other more complex emotion based factors such as Trust, Credibility and Loyalty. In addition to measuring these factors, four questions in the SUPR-Q were constructed to measure users’ perceptions of usability factors. The SUPR-Q displays a strong correlation with a SUS score, $r = .96$. $p < .001$, meaning these four questions account for 93% of the variation in SUS (.96 squared).

These four usability questions; ease of use, enjoyment, findability and ease of navigation correlate with the SUS and account for about 93% of the variation in SUS scores (Sauro, 2010). A factor analysis was used to determine the questions that cluster together and create the subscales that make up the SUPR-Q. Most of the questions in the SUPR-Q that are attitudinal correlate with the SUS, but four usability questions (the average and the sum of the four “usability” questions) correlate almost exactly with the SUS. The additional nine questions provide information that the SUS does not. These questions make up the Trust, Appearance, Value and Comfort scores of the metric (Sauro, 2010). A psychometric questionnaire tool should provide both a valid and reliable instrument as well as a relative ranking based on a dataset. While the SUS is a valuable instrument, a raw SUS score is difficult to interpret without knowing how other websites score (e.g., is
a 70 good?). Questionnaires like the WAMMI and SUPR-Q both provide relative rankings expressed as percentages. Therefore a percentile score of 50 is an average score, or roughly half the websites have received better scores and half received worse scores. However, one disadvantage of the WAMMI is that no information is provided regarding what websites make up the scoring system. In any large dataset there will be some websites that are good comparisons and others which are irrelevant. The SUPR-Q provides specific information regarding what websites are in the dataset. Additionally, SUPR-Q scores can be compared to a subset of up to 10 other websites or only websites with a particular category. Data from approximately 112 websites make up the SUPR-Q database. Most websites are US based and were tested on users from North America. Roughly 10% of the users come from countries outside of the US & Canada including: UK, Australia, Romania, India, Japan, Mexico and Hong Kong. A partial list of the websites in the database is shown in the table below (see Figure 1.5). They include large and small ecommerce websites, mobile phone carriers, travel and government websites (Sauro, 2010).
<table>
<thead>
<tr>
<th>Retail</th>
<th>Airlines</th>
<th>Floral Service</th>
<th>News &amp; Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Amazon</td>
<td>- United Airlines</td>
<td>- FTD</td>
<td>- IMDB</td>
</tr>
<tr>
<td>- eBay</td>
<td>- Frontier Airlines</td>
<td>- 1800 Flowers</td>
<td>- Yelp</td>
</tr>
<tr>
<td>- Zappos</td>
<td>- American Airlines</td>
<td>- ProFlowers</td>
<td>- Craigslist</td>
</tr>
<tr>
<td>- Apple</td>
<td>- Southwest</td>
<td></td>
<td>- Motley Fool</td>
</tr>
<tr>
<td>- Target</td>
<td>- Jet Blue</td>
<td></td>
<td>- Yahoo</td>
</tr>
<tr>
<td>- JC Pennys</td>
<td>- Delta</td>
<td></td>
<td>- Wall Street Journal</td>
</tr>
<tr>
<td>- Walmart</td>
<td>- Frontier</td>
<td></td>
<td>- CNN</td>
</tr>
<tr>
<td>- Hermes</td>
<td>- Boston Proper</td>
<td></td>
<td>- FoxNews</td>
</tr>
<tr>
<td>- Container Store</td>
<td>- Crate &amp; Barrel</td>
<td></td>
<td>- New York Times</td>
</tr>
<tr>
<td>- Crumpler</td>
<td>- Oriental Trading</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Sears</td>
<td>- LL-Bean</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Michaels</td>
<td>- Payless Shoes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Pier 1 Imports</td>
<td>- Sears</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Macys</td>
<td>- LL-Bean</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Netflix</td>
<td>- Michaels</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- BestBuy</td>
<td>- Payless Shoes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- OfficeMax</td>
<td>- Pier 1 Imports</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- OfficeDepot</td>
<td>- Macys</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3rd Party Travel</th>
<th>3rd Party Automotive</th>
<th>Financial Services</th>
<th>Social Networking</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Travelocity</td>
<td>- Edmunds</td>
<td>- PayPal</td>
<td>- Facebook</td>
</tr>
<tr>
<td>- Expedia</td>
<td>- Kelly Blue Book</td>
<td>- Fidelity</td>
<td>- Twitter</td>
</tr>
<tr>
<td>- Orbitz</td>
<td>- Cars</td>
<td>- TD Ameritrade</td>
<td>- LinkedIn</td>
</tr>
<tr>
<td>- Kayak</td>
<td>- Autotrader</td>
<td>- Vanguard</td>
<td>- Flickr</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Government</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>- Illinois</td>
<td>- 3rd Party Automotive</td>
<td>- Financial Services</td>
<td></td>
</tr>
<tr>
<td>- Colorado</td>
<td>- 3rd Party Travel</td>
<td>- Government</td>
<td></td>
</tr>
<tr>
<td>- California</td>
<td>- Retail</td>
<td>- Airlines</td>
<td></td>
</tr>
<tr>
<td>- CDC</td>
<td>- Airlines</td>
<td>- Floral Service</td>
<td></td>
</tr>
<tr>
<td>- USA.gov</td>
<td>- Retail</td>
<td>- News &amp; Information</td>
<td></td>
</tr>
<tr>
<td>- New York State</td>
<td>- Airlines</td>
<td>- Retail</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1.5: Partial list of websites in the SUPR-Q database
Using the SUPR-Q for the Kelley Blue Book Desirability Study

Method

The Kelley Blue Book Desirability study was conducted remotely using Keynotes’ WebEffective (a remote usability tool) software. To limit fatigue and support participant completion, each session was designed to take approximately twenty to twenty-five minutes. Instructions were standardized and presented to all participants at the beginning of the study to ensure the validity of the test (see Figure 1.6).

Figure 1.6: Instructions presented to participants prior to viewing each Concept A and B
Participants were presented with three unbranded key pages for two potential designs. The Kelley Blue Book creative team hypothesized that participants would experience Design “A” with its softer edges, warmer more saturated colors, conversational tone and more geometric font as the more “approachable” concept (see Figure 1.8). Conversely, the designers assumed Design “B”, with its harder edges; cooler more monochromatic color palette, directional and “matter of fact” copy tone and more traditional font would be experienced as the more “corporate” concept (see Figure 1.9). Participants were encouraged to explore the set of key screen shots for each concept by clicking, scrolling and interacting with a set of limited functional elements (all scroll bars, global navigation drop downs, show-hide panel/accordions and hovers were actionable on each of the three prototype pages) (see Figures 2.0 and 2.1). The order of designs “A” and “B” was counterbalanced, although the order of the key page screen shots was consistent.
Figure 1.8: Design “A” The “Approachable” Concept (hypothesized favorite)
Figure 1.9: Design “B” The “Corporate” Concept (hypothesized least favorite)
Functional interactions (e.g., show-hide panel). In “show” state.

Functional drop down menus

Functional scroll bar

Navigation forward and backward through 3 key pages.

Figure 2.0: Design “A” (Page 2) The “Approachable” Concept (Prototype Functionality)

Figure 2.1: Design “B” (Page 2) The “Corporate” Concept (Prototype Functionality)
Part I of the Kelley Blue Book Desirability study was based on a modified version of the Microsoft Desirability Study product reaction card experiment design originally developed by Joey Benedek and Trish Miner for Microsoft in 2002. This original study design was developed to address methods for assessing how users perceive different potential designs by having them select from a list of 118 attributes they associate with each design. For the Kelley Blue Book Desirability study, participants chose three words from a provided list of 50 attributes for both concepts “A” and “B”. The study maintained the required ratio of 40% negative/neutral to 60% positive words recommended from the Microsoft experiment design. The list of 50 attributes for the Kelley Blue Book study consisted of a balance of 20 negative/neutral words and 30 positive words. The fifty word set contained seven brand attributes that the design team wanted to discover if participants selected to describe their hypothesized preferred design. Revisions to the Microsoft Product Reaction Card study design introduced in 2004 in a paper by Anderson, de los Reyes, Kelly, et al. included the addition of statements containing key brand attributes. Part II of the Kelley Blue Book study incorporated this addition into its experiment design. Participants were presented with seven statements about each design they had just viewed. They were then asked to rank the degree to which they agreed with the following statements on a 1-5 Likert-type scale.

- This website made me feel **empowered**
- This website was **helpful**
- This website was **trustworthy**
- The information on this website seemed **accurate**
- This website was **educational**
- This website seemed **approachable**
- This website was **modern**
Each of the seven statements contained one of the attributes the Kelley Blue Book design teams postulated their preferred design embodied (i.e., empowered, helpful, trustworthy, accurate, educational, approachable and modern).

In Part III of the study, participants were asked to complete the psychometric SUPR-Q tool as a post-task questionnaire. In Part IV of the study, participants were presented with three optional questions about the design (best thing about the design, worst thing about the design and general comments). Concepts remained viewable by participants throughout parts one through four of the study. Participants repeated the same steps for each design before being presented with Part V of the study. In Part V of the study, small screen shots of each unbranded design (design “A” and design “B”) were presented on the same page (counter balanced in order). Participants were asked to choose their favorite from the two designs and provide comments explaining their preference.

Participants

Approximately 200 participants were recruited for the unmoderated remote online Kelley Blue Book Desirability study. During analysis, 20 participants were eliminated due to questionable responses on a “speeder” detection question. “Speeders” were defined for the study as participants who attempted to rush through the study by spending less than 30 seconds on any of the pages of the study or arbitrarily clicking on questionnaire answers. Participants were also eliminated if non-optional answers included “junk” responses (i.e., a nonsensical repetition of letters or numbers). The final count of participants available for data analysis was 180.

A fifty-five percent male and forty-five percent female mix of U.S. based participants between the ages of 18 - 65 were recruited via an external recruitment company. This
balance of gender and age demographics was representative of the average Kelley Blue Book audience. Average demographic web trend data for KBB.com was provided for the study from the Kelley Blue Book analytics department for the fourth quarter of 2010. In addition to age, gender, education, internet usage and household income questions, potential participants from the external recruiter’s database were asked to complete a special car purchase behavior screener. This screener was used to identify participants who had purchased a car within the last year or were in-market to purchase a car within the next year. Qualifying participants from both the “recent purchase within the last year” and “intends to purchase within a year” categories were then required to complete an additional questionnaire to segment them into personas.

Based on previously conducted ethnographic research conducted by Kelley Blue Book user experience researchers, three “persona determining” questions and answers were developed:

1. Which best describes the way you do research on which car to buy?
2. Which sources do you use to do research on which car to buy?
3. Which best describes the way you decide on a BRAND (i.e., Ford or Honda) of car to buy?

The answers to each of the three question (i.e., I like to look at all the options and get all the information possible, even if it takes longer to make a decision) were associated with a specific persona (i.e., Persona #3). Participants were instructed to select the answer for each question that most closely matched their perspective in relation to the car shopping process.

The following criteria were constructed to allow participant segmentation into the appropriate personas based on participant’s answers:
• If the answers to all three questions match one specific persona, then a participant is assigned to that persona.

• If answers from any two questions match one specific persona, then a participant is assigned to that persona

• If answers from all three questions do not match one specific persona then reject that respondent as an eligible participant for the study

A total of 30 participants were targeted for each of the six (6) personas. An exact match of the demographic screener used by the third party recruiter was included to be answered by participants post-test. This allowed Kelley Blue Book researchers to cross correlate demographic answers pre and post-test.

Analysis

For the Kelley Blue Book Desirability study, the SUPR-Q was used as a post-test questionnaire. After filtering for incomplete or suspect data, a total of 180 participants in the remote study had correctly completed the 13-item questionnaire after viewing each of the two design versions provided. Twenty respondents who were identified as “cheaters” or “speeders” were excluded from the final analysis. “Speeder” participants were flagged in the remote usability research Webeffective tool if they spent less than 30 seconds on any of the key pages or answering any questions in the study. “Cheater” participants were defined as any participant who answered questions with “junk” (i.e., repetitive and/or nonsense numbers or letters) responses in mandatory answer fields. A “cheater trap” question was also included. Participants could answer question number three (i.e., I enjoy using this website) positively with a rating of three or more, or negatively with a rating of three or less. The “Cheater trap” was the opposing sentiment, question number
nine (i.e., I don’t enjoy using this website). If participants provided the same rating for both questions (other than neutral 3) they were flagged as a “cheater”.

A minimum of 150 total participants was required to appropriately reflect a year-over-year auto buying population of approximately fifteen million. To reduce sampling error to between 15-17% when conducting persona specific analysis, for each of the six personas, a minimum of 20 participants was targeted. Once the initial raw data was filtered, four of the six persona based segments contained at least 30 participants and two contained at least 25 participants. (A sample size of 80 participants per segment would have been required to reduce sampling error to ten percent for a 90% confidence interval analysis. However, a smaller sample size was decided upon to limit additional recruitment time and costs associated with the larger sample size.)

A multiple regression comparison was conducted for each metric analyzed. The multiple regression analysis (a multivariate technique) was used to look at the correlation between the dependent variable (likelihood to recommend) and the independent variables (i.e., Usability, Loyalty, Trust/Value/Comfort/Credibility and Appearance). A multiple regression analysis takes into account the inter-correlations between independent variables. The remaining correlation (called a beta coefficient) measures how much each variable contributes to the Net Promoter. All data were analyzed for ninety percent confidence intervals (per persona confidence intervals where at the 85% CI). Scored responses to the SUPR-Q questions were to provide an overall raw and standardized picture of Usability, Loyalty, Credibility (Trust/Value/Comfort) and Appearance.
Based on the participant’s responses, there was no statistically significant difference between design “B” the “corporate” design, with a percentile rank of 80% and design “A” the hypothesized preferred design, with a percentile rank of 76%, (p > .4) (see Figure 2.2).

Figure 2.2: SUPR-Q Results for the Kelley Blue Book Desirability Study

SUPR-Q questionnaire responses measuring Credibility (Trust/Value/Comfort), Usability and Appearance and were not statistically significant between designs “A” and “B” (Credibility (Trust/Value/Comfort) = p > .9, Usability = p > .75) However, responses measuring Loyalty were statistically significant (p < .04) in favor of design B, the more “corporate” concept (see Figure 2.3).
When comparing the proposed Kelley Blue Book redesigns to third party automotive website competitors in the SUPR-Q database, both versions “A” and “B” scored higher on the SUPR-Q than Competitor #1 and marginally better than Competitor #2 and Competitor #3 (see Figure 2.4).
Both redesigned concepts scored statistically higher for Usability than the current KBB.com site and all competitors p < .01 (see Figure 2.5).

Figure 2.4: SUPR-Q overall and sub-scale scores

Figure 2.5: SUPR-Q Usability Raw Scores For Designs “A”, “B” and Competitors
There were no statistically significant differences between any versions for the Credibility (Trust/Value/Comfort) SUPR-Q scores (see Figure 2.6).

For the Loyalty score, all three KBB designs were statistically higher than the competitors (p <.01) but not different from each other (see Figure 2.7).
Design concepts “A” and “B” were statistically higher than all other versions for Appearance scores but not statistically difference than each other (see Figure 2.8).

Figure 2.8: SUPR-Q Appearance Scores for Designs “A”, “B” and Competitors
There were no statistically significant differences in SUPR-Q scores between Concept “A” and “B” across personas (see Figure 2.9).

![Figure 2.9: SUPR-Q Average Scores for Designs “A”, “B” by Persona](image)

Discussion

In an effort to quantitatively measure the components of “Desirability” across target audiences for two proposed concepts of a website redesign, the Kelley Blue Book study utilized several methods. These methods included a combination of Microsoft’s Product Reaction Cards and a recently introduced post-test psychometric measurement tool, the SUPR-Q. The Product Reaction Cards provided measurable terminology for study respondents to describe their emotion based attitudes and perceptions of the proposed design concepts. The SUPR-Q questionnaire attempted to provide quantifiable measurement of study participant’s unobserved but self-reported perceptions of Credibility (Trust/Value/Comfort), Loyalty, Usability and Appearance related to the proposed redesign concepts. Additionally, the SUPR-Q percentile rank and database of
other 3\textsuperscript{rd} party websites were intended to provide the ability to compare the proposed design concepts’ rankings against other websites in the same category. For the Kelley Blue Book Desirability study, the SUPR-Q revealed to the design team that both proposed redesigns (concepts “A” and “B”) scored higher in perceived Usability than the current live site and all other 3\textsuperscript{rd} party auto site competitors in the SUPR-Q database.

In contrast to the SUPR-Q, the more often used, SUS (System Usability Scale) rating may provide a picture of actual or perceived overall usability of a software product or website. However, when the objectives of the research are more emotion based, as in a Desirability study, the SUS alone is an inadequate measurement tool. Desirability contrasts with Usability and Usefulness/Utility in that the typical usability/usefulness study answers the central question, “Can the product be used, and does it adequately meet a need?” In a Desirability study, the objective is to answer multiple questions such as, “Do I want to use the website/product?”, “Is it aesthetically pleasing”, “Does it evoke a positive emotional response?” and further “Does it move toward a company’s target brand attributes?” (Benedek et al., 2002). For the Kelley Blue Book design team, the SUPR-Q allowed a comparison of a combination of emotion based factors, including perceived usability, related to a set of proposed designs. Perhaps, most importantly, it also revealed that the creative team’s preferred concept did not emotionally resonate with participants across personas, retention or acquisitional audiences. Nor, was it significantly more emotionally effective than an assumed less desirable design. The results of the Microsoft Product Reaction cards questionnaire, Part I of the Kelley Blue Book Desirability study, helped to support data triangulation. In an analysis of the frequency distribution of the positive and negative attributes provided to respondents to
describe both designs “A” and “B”, both designs scored almost equally in positive sentiment selection. Similarly, the SUPR-Q results demonstrated no statistically significant preference by participants overall or even across persona’s to concept “A” over concept “B”. It may be significant to note, that this lack of any clear differentiation in participants preferences between the designs may be attributed to the similar interactions, color schemes and copy treatments in the two designs.

With the added benefit of the SUPR-Q percentile rank, a comparison to other sites in the 3rd party auto site category was possible, providing an informative point of reference for the design team. For example, proposed concepts “A” and “B” scored similar to one another on Appearance and Loyalty. However, the SUPR-Q percentile rank provided the ability for comparisons to other 3rd party automotive websites in the database. This capability allowed the team to note that both designs scored higher in Appearance and Loyalty than other SUPR-Q database websites in the same category, including the KBB.com website design that was live at the time.

Although the SUPR-Q questionnaire provided several invaluable insights for the Kelley Blue Book design team, many open questions remain. First, the primary objective of a Desirability study is to ascertain the impact of the visual design on a users’ perception of a product, website or hand held device. Previous research does support the assertion that a product or website’s visual design has a significant impact on a users’ perception of Credibility, Trust and Usability:

There is a significant amount of research that supports the view that design, aesthetics and usability are inexorably linked (Alsudani & Casey, 2009; Fogg et al, 2003; Lindgaard, Fernandes, Dudek, & Brown, 2006; Norman, 2002; 2004; Tractinsky, Katz & Ikar, 2000; Tractinsky, 2006; Zhang, 2009). But how are they
linked? Fogg et al. (2003) found that over 45% of consumers made judgments about the credibility of websites based on the site design, “including layout, typography, font size, and color scheme (p. 5).” Studies have found that users make these judgments about visual stimulus in a very short amount of time (Alsudani & Casey, 2009). Robins and Holmes (2008) found that subjects judge the credibility of the content of a website based on its appearance in 3.42 seconds. Several studies have demonstrated that aesthetic judgments also affect a system’s perceived usability (Mbipom and Harper, 2009; Tractinsky, Katz & Ikar, 2000). These studies have demonstrated that from the user’s perspective “what is beautiful is usable” (Tractinsky et al, 2000, p.129), and this correlation between aesthetics and usability remained strong in post-use evaluation. (David and Glore, 2010).

Loyalty, has also been linked to aesthetics in previous research:

While early electronic commerce research indicated that website design elements could influence online satisfaction and sales (Loshe, Spiller 1999), most research ignored the importance of website design factors and how they can influence customer satisfaction and the likelihood that customers will revisit a website resulting in customer loyalty (Fornell, Krishnan, Mithas,Ramasubbu 2003). Online loyalty or e-loyalty has been defined as a consumer’s intention to buy from a website or to visit it again (Bowes, Bonanni, Cyr, Ilsever, 2005; Flavian, Guinaliu, Gurrea 2005; Yoon, 2002). Studies on the Internet suggest that if users are satisfied with the design of a website, they are more likely to exhibit e-loyal behavior (Anderson, Srivanan, 2003) (Cyr, Head, Ivanov, 2006).

However, while perceived Credibility, Loyalty, Appearance and Usability may contribute to a products “Desirability” these factors may not be all inclusive. For example, an appropriate post-Desirability study questionnaire might additionally include questions and/or statements related to (when appropriate) factors such as perceived Playfulness (Fun and Delight) or perceived Entertainment (Enjoyment) (Cyr, Head, Ivanov, 2006). When considering hand held devices, previous research suggest factors such as perceived Risks, Costs and Compatibility with mobile commerce may also contribute to “Desirability” (Wang, Wu, 2004).

Additionally, while it can be argued that the four factors of the SUPR-Q are appropriate to measure a product’s Desirability, the specific questions and wording in the SUPR-Q may not be. Statements in the Credibility (Trust, Value and Comfort) category
of the SUPR-Q potentially imply a participant has had a more robust experience with a site (i.e., I feel comfortable purchasing from this website, I feel confident conducting business with this website, etc.). This perception would be based almost exclusively on visual aspects with limited interactions or navigation of the site. For the most part, the low or even medium fidelity concepts often leveraged in Desirability studies do not offer those rich interactions and experiences. Similarly, the Usability related questions of the SUPR-Q, (an area in which both proposed concepts “A” and “B” scored higher than competitors and the current Kelley Blue Book website), also imply participants have potentially navigated or completed a task on the site sufficient enough to effectively answer the questions (i.e., This website is easy to use, I am able to find what I need quickly on this website, etc.).

Further, although the Kelley Blue Book Desirability study was unbranded and all attempts were made to conceal any brand affiliation, the number of 3rd party auto sites in the market is limited. Participants may have imposed any of their more positive experiences on the proposed designs, therefore leading to more positive usability scores. Additionally, although both concepts were limited in functionality (participants could not start or complete a full task with the prototypes) they were “finished” and professional looking. In the absence of any definitive experiences to the contrary, participants may have assumed more positive experiences for ease of use and navigation. In contrast, the fully functioning live competitive 3rd party auto sites in the SUPR-Q database (with any potential navigation, way finding or site technology errors that implies) may have fared less well comparatively.
The constraints of the remote user research method employed for the Kelley Blue Book Desirability study (despite “speeder” and “cheater” detection efforts noted earlier), make it difficult to determine the exact time participants spent perusing the limited functionality of each design concept. Therefore, the degree of thoroughness related to participant’s investigation of each concept cannot be definitively determined.

The SUPR-Q is a recently developed psychometric questionnaire and other research using it as a tool to specifically measure website Desirability is limited. Currently, the only other comparative research leveraging the SUPR-Q as an effective psychometric measurement tool has been conducted by the tool’s creator, Jeff Sauro. To date, no other independent studies have been conducted measuring the validity and reliability of the questionnaire. Further research comparing and contrasting its use as a post-test questionnaire to appropriately measure website design Desirability would need to be conducted to explore its efficacy and effectiveness.

To this end, a future study for the Kelley Blue Book User Experience Research team might well include a post redesign launch of the SUPR-Q on the live KBB.com website. In this manner, SUPR-Q results from the concept chosen based on the Desirability study could be compared to results from the redesigned live site as well as competitors in the 3rd party auto-site category of the SUPR-Q database. A comparative analysis of the SUPR-Q per category results (Credibility, Loyalty, Appearance and Usability) pre and post redesign launch could support a more in-depth investigation of the tool as an effective Desirability study instrument.

With the exception of the Microsoft studies in 2002 and 2004 that developed the Microsoft Product Reaction Cards method, a limited number of studies have been
conducted with the intention of exploring methods to measure website and product desirability. Fewer still include psychometric questionnaires in the experiment design. In a Delgado-Ballester et al. 2003 study, “Development and Validation of a Brand Trust Scale,” the focus was exclusively on developing a Likert-type tool to effectively measure consumer trust of the brand. The desirability evoked by a website or a product’s design alone as opposed to the desirability evoked by the brand was not considered. Another study was conducted in 2010 by Edward DeGuzman and Julie Schiller from Autodesk, entitled, “How Does This Look? Desirability Methods for Evaluation Visual Design.” This study did focus on examining desirability of a new AutoCAD software product by leveraging a modified version of the Microsoft Product Reaction cards method. However, unlike the Kelley Blue Book Desirability study, the DeGuzman and Schiller study did not attempt to explore desirability using categories of perceived Credibility, Appearance Loyalty or Usability or even augment their study design with a psychometric questionnaire.

Similarly, an article appearing in the 2011 issue of the Usability Professional Association Magazine authored by Hilary Palmen entitled, “How Was It For You?, Helping People Describe Their Experiences”, also leveraged a modified version of the Microsoft Product Reaction Card’s method. However, the experimental design discussed focused on Desirability by providing a vocabulary for participants to describe their experiences with a product but did not specifically focus on how well the selected terms reflected users’ perceptions by any SUPR-Q factors (i.e., Credibility, Loyalty, Appearance and Usability). A 2010 study by Barnum and Palmer, built upon the Microsoft Product Reaction card method to help study participants tell a rich story about their experiences with several different products and services. The authors additionally
noted that the cards “should not be used as the sole means of getting participants’ feedback regarding their experience. They work best when used along with other satisfaction survey instruments” (Barnum and Palmer, 2010). Unfortunately, the authors did not provide specific detailed information about questionnaires used in the study. Neither, did they expound on what specific information they hoped to glean from a post-test satisfaction instrument that the Microsoft Product Reaction Card method alone did not provide.

Based on the literature regarding Desirability studies conducted in the past several years, Microsoft’s Product Reaction Cards seemed to have become an integral aspect. Further development of a tool to measure the multi-dimensions of Desirability might well address not only the frequency of word choices in a Product Reaction Card based method but also the strength of any of the key words in the selection set. Keywords could be pre-categorized by factors that correlate to questionnaire categories, (as in the SUPR-Q’s Credibility, Loyalty, Appearance and Usability factors). This pre-categorized selection of keywords, along with their associated strength might provide valuable and rich supportive information if analyzed in relation to questionnaire data.

For the Kelley Blue Book design and research teams, many valuable lessons were learned from the execution of the company’s first Desirability study. Perhaps the most important being that the subtle visual differences in colors, interactions and whitespace between the two proposed design concepts resulted in participants being unable to clearly define a concept “winner”. Future Desirability studies undertaken by the Kelley Blue Book User Experience team may be more effective if preliminary “mini” in-lab Desirability research is conducted with a smaller number of participants. Participants could leverage the Microsoft Product Reaction Cards to verbalize their thoughts and
feelings about the concepts. However, this first step might include presenting a wide variety of several distinctly different design concepts (perhaps 3-7) to participants. In this way, user experience researchers might then be better equipped to aid designers in refining to a more differentiated concept selection set for a multi-method remote study with a larger sample size.

Finally, attempting to gauge users’ emotional responses to a design is an extremely complex endeavor. Users’ experiences of visual designs are subjective and multifaceted and myriad of design aspects can impact a user’s response to a website, product or handheld device. While research methods related to adequately measuring Desirability are still in their nascent stage, perhaps comments from a posting to Jakob Nielsen’s online website Alertbox, in July of 2002 expressed the importance of such research best:

Usability will not save an undesirable product, site or service. In a world of plenty, demand is the scarcest resource of all. Attention and desire are the new scarce resources in an era of unprecedented product availability and technical knowhow (Nielsen, 2002).
References


Dzida, W., Wiethoff, M., and Arnold, A.A., (1993), ERGOguide: the Quality Assurance Guide to Ergonomic Software. Delft University of Technology, Dept. of Work and Organizational Psychology, PO Box 5050, 2600 GB Delft, the Netherlands


