Beyond Usability: Evaluating Emotional Response as an Integral Part of the User Experience

Anshu Agarwal
User Experience
Salesforce.com
San Francisco, CA
aagarwal@salesforce.com

Andrew Meyer
Stanford University
Palo Alto, CA
almeyer@stanford.edu

Abstract

The role of emotion as an integral component of user experience has mostly been overlooked in the HCI literature. Instead, usability has been relied upon as the key indicator of user experience. We developed a methodology that combined verbal and nonverbal emotion scales. A usability study was then conducted, in which we collected both traditional usability metrics and emotional response data. Results indicated insignificant differences in usability metrics but numerous significant differences between emotional responses of users. Exploration of these emotional responses successfully provided additional insight into the user experience.

CHI 2009, April 3 – April 9, 2009, Boston, MA, USA
ACM 978-1-60558-247-4/08/04.

Keywords

Usability, emotion, user experience, user interface, design, metrics, methodology

ACM Classification Keywords

H.5.2 User Interfaces: Evaluation/methodology, theory and methods, user-centered design.

Introduction

Much of what we do as humans is driven by our emotions. Many psychologists argue that it is impossible to act or think without engaging, at least unconsciously, our emotions [17]. While research concludes that emotion is a fundamental component of being human, the HCI community -- a community which studies the interaction between humans and computers -- has mostly overlooked emotion as a component of user experience. Recently, the topic of emotion in HCI has attracted increased research attention [2]. Currently, most publications discuss emotion metrics and their application at a more theoretical level [1, 2] or these metrics are applied in fields other than HCI [5, 6, 7, 9, 15], such as ergonomics and aesthetics. Very few case studies or real world examples examining emotion in an HCI context have been published.

A review of the multidisciplinary literature shows the significant impact emotion may have on issues central to the HCI community and to disciplines beyond HCI. Studies have shown that humans are more efficient and creative problem solvers when they are happy [11]. Research also indicates that emotion is closely tied to user acceptance and satisfaction [9], may heavily impact purchase intention [9], and often serves as a primary motivation for consumptive behavior [10]. These findings are bolstered by de Martino's 2006 brain-imaging study, which found that certain brain regions associated with emotion are highly active when people make buying decisions [4].

In this paper, we briefly review existing emotion measurement tools. We then discuss the development of a method that combines both verbal and nonverbal scales to assess emotional responses. A usability study of two interfaces was then conducted. Ultimately, this paper discusses the significant value emotion measurement provides to better understand the complete user experience. We conclude with a discussion and opportunities for future research in the realm of HCI and emotion.

What is Emotion?

Before we continue onto emotion measurement, we first need to define what we mean by "emotion." Although we often use the term "emotion" in our daily lives, coming up with a precise and scientifically respectable definition of the term "emotion" is notoriously difficult. As one might imagine, there are many definitions of "emotion" in the relevant literature [14]. Nevertheless, there are two generally agreed on aspects of what actually constitutes human emotion [2]. First, emotion is a psychological reaction to events relevant to the needs, goals, or concerns of an individual. Second, emotion is comprised of physiological, affective, behavioral, and cognitive components [2]. In this paper, our interpretation of "emotional response" is often technically termed sentiment, the assigned emotional property of an object such as an interface.

Measuring Emotion

Emotion is an inherently complex construct to study. Humans have trouble describing how they feel and distinguishing between different emotions [5, 6]. It is also difficult to pinpoint the exact cause of a particular emotion, partially because emotions can change instantaneously [2]. In an attempt to meet these challenges, researchers have created many different emotion measurement tools. These tools include verbal measurement tools, nonverbal measurement tools, and physiological measurement tools. This paper will only explore the application of verbal and nonverbal measurement tools because our study was conducted remotely, which prevented us from exploring physiological responses as indicators of emotional response.

Verbal Measurement Tools

Verbal emotion measures have been developed and utilized primarily in marketing and advertising research. These tools usually take the form of self-reports where respondents use a scale to record their emotions. The main criticism of verbal techniques is that they only capture respondents' conscious emotional states. Additionally, in order to effectively assess emotional response, comprehensive verbal measures tend to be lengthy. Since emotions are immediate and automatic, the extensive time it takes to complete a verbal measure may distort how users initially felt. Emotions are also fleeting and hard to distinguish, so respondents may have difficulty remembering how they felt only minutes earlier. Researchers have developed continuous self-report measurements to address the problem of the fleetingness of emotion; however, these tools are best applicable in studies in which the subject passively participates in tasks. Studies requiring active participation, such as use of an interface, may often preclude the subject from continuously using devices (i.e. physical dials or onscreen sliders) traditionally used in continuous self-reporting methods.

Finally, verbal measures are language-dependent, precluding their application to certain populations and children. Despite these limitations, verbal measurement tools remain popular for emotional assessment because they are relatively easy to develop and use.

Likert scales and Semantic Differential scales are the two most prominent types of verbal scales. Respondents must choose along a scale to reflect how they feel at any given moment. While Likert scales typically rate one item from strongly disagree to strongly agree, Semantic Differential scales place bipolar adjective pairs at each end of the scale. Respondents select anywhere in-between the two adjectives on the scale to respond.

The following are a few examples of verbal emotion measurements: Standardized Emotional Profile [12], Feelings Toward Ad scales [7], Leavitt Reaction Profile [16], and the PAD Semantic Differential Scale [17].

Nonverbal Measurement Tools

It is challenging to find an effective and valid nonverbal measure of emotion. The most common nonverbal measures usually include visual representations of emotion that participants select to characterize how they feel. Human-like depictions of emotion, such as a smiling face, have been validated cross-culturally as consistently interpreted. These human-like visual representations may therefore be used for a wider array of respondents compared to verbal measures. Nonverbal scales also aim to capture unconscious emotional responses and incorporate a certain amount of 'fuzziness' appropriate to the study of emotion.

The following nonverbal measures of emotion tend to be the most recognized in academic research: PrEmo [5], Self Assessment Manikin [1], Facial Action Coding System [8], and Emocards [6].

Method

Study Measure Development

Our goal was to use an emotion measure that could be low investment, quick, easily understood, and incorporated into a traditional usability test. Since our study was conducted remotely, we disregarded physiological measurements and relied on methods which could be deployed in this fashion. While numerous measures existed from previous research, it was apparent that no one emotion measure alone would suffice as a reliable assessment of emotion. Although ‘fuzzy’ nonverbal measures seemed more appropriate to assess emotion, most of these measures were all very experimental and of unknown validity. We decided to combine an extensively used verbal scale with a more experimental nonverbal emotion measure to improve the strength of our emotion methodology.

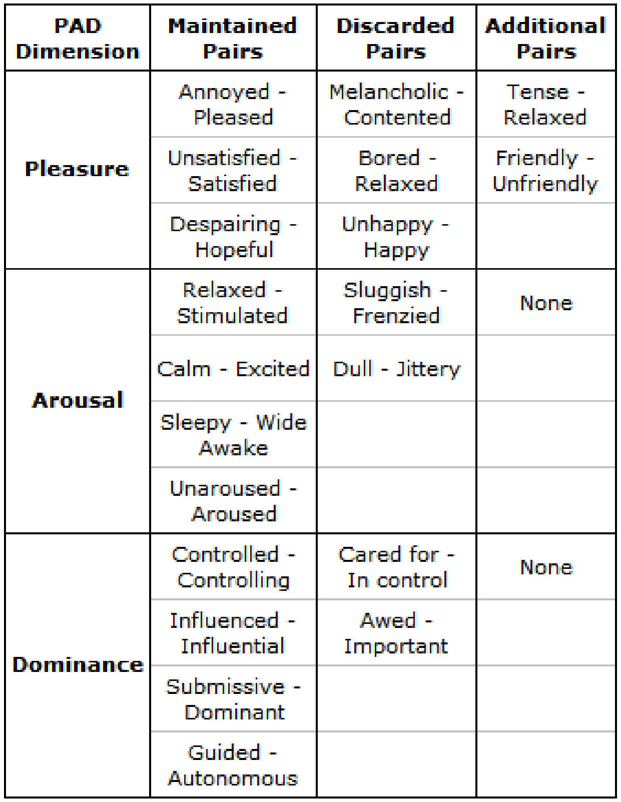
For the verbal component, we chose to pursue the PAD Semantic Differential Scale (PAD scale) developed by Mehrabian and Russell [17]. This scale included a set of bipolar adjective pairs that were rated along a nine point scale. Based on results of their original study, the PAD scale was shown to measure three important aspects of emotion: Pleasure, Arousal, and Dominance. Pleasure may be defined as a positive affective state which is separate from feelings such as preference and reinforcement [17]. Arousal refers to an emotional state from sleepy to very excited [17]. The final dimension, Dominance, refers to the extent to which a person feels unrestricted or free from outside control [17].

The PAD scale is one of the most extensively used and validated tools for measuring emotional response. Although it has been empirically supported in numerous studies related to advertising and marketing, its validity in application to software interfaces was undetermined.

We reviewed Mehrabian and Russell's original adjective sets to ensure that the pairs were relevant to interface emotional responses. We also considered the length of Mehrabian and Russell's scale and aimed to reduce the number of items. Based on this review, we added in two additional bipolar adjective pairs for the dimension of Pleasure: Tense-Relaxed and Unfriendly-Friendly. We also eliminated seven adjective pairs which appeared less relevant to software evaluations: happy-unhappy, melancholic-contented, bored-relaxed, sluggish-frenzied, dull-jittery, cared for-in control, and awed-important. These seven pairs were removed because they would not make sense in the context of evaluating a software interface. These changes can be seen in table 1, which displays the bipolar adjective pairs that we maintained from the original set, those that we discarded from the original set, and the adjective pairs that we added into the Semantic Differential scale.

We selected the Emocard tool by Desmet for the non-verbal component (figure 1) [6]. The Emocard tool consists of sixteen cartoon-like faces, half male and half female, each representing distinct emotions. Each face represents a combination of two emotion dimensions, Pleasure and Arousal, which are identical to two of the dimensions included in the PAD scale. Based on these dimensions, the Emocards can be divided into four quadrants: Calm-Pleasant, Calm-Unpleasant, Excited-Pleasant, and Excited-Unpleasant. In the study, user reactions that were more pleasant and higher in arousal were considered desirable. Results in the Calm-Pleasant and Excited-Pleasant quadrants were therefore interpreted as positive results.

Both the PAD scale and the Emocards were placed into an online survey to be used in our study. PAD scale items and Emocard images were presented in randomized order.



PAD Dimension	Maintained Pairs	Discarded Pairs	Additional Pairs
Pleasure	Annoyed - Pleased	Melancholic - Contented	Tense - Relaxed
	Unsatisfied - Satisfied	Bored - Relaxed	Friendly - Unfriendly
	Despairing - Hopeful	Unhappy - Happy	
Arousal	Relaxed - Stimulated	Sluggish - Frenzied	None
	Calm - Excited	Dull - Jittery	
	Sleepy - Wide Awake		
	Unaroused - Aroused		
Dominance	Controlled - Controlling	Cared for - In control	None
	Influenced - Influential	Awed - Important	
	Submissive - Dominant		
	Guided - Autonomous		

table 1. *We selected the PAD Scale for our verbal scale. Although we maintained most of the original adjective word pairings, we discarded a few and then added in additional pairings to ensure the scale was concise and relevant to software interfaces.*

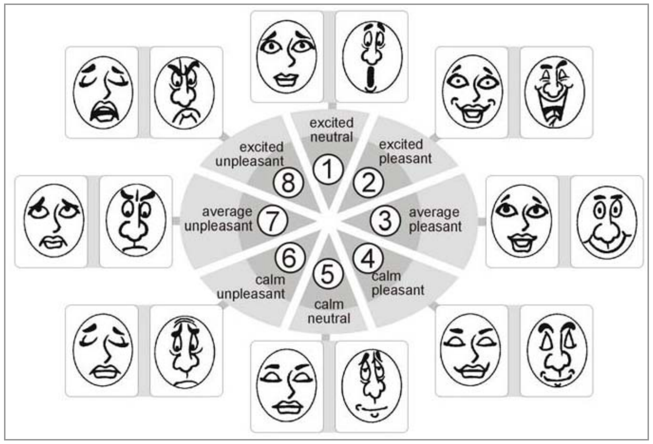


figure 1. The Emocard tool was an effective nonverbal measurement of emotional response which used human-like representations of emotion.

Study Procedure

A between-subjects study was then conducted to test our new combined emotion measure with users. We decided to run a comparative study between two versions of a Customer Relationship Management (CRM) application interface using traditional usability measures (time on task and number of errors) as well as the new emotion measure we constructed. The use of emotion measure allowed us to see if there existed differences between the two interfaces that were not discoverable with the standard usability metrics.

Twenty-two participants, thirteen male and nine female, were assigned to assess one of two CRM interfaces. Assignment ensured that participants had not had any prior direct experience with the interface they were evaluating in the study. Seven comparable tasks between the two interfaces were created. Abbreviated versions of these tasks may be seen below.

Task 1: Find and read help information on how to create Leads.

Task 2: Manipulate the calendar entry ‘Visit Bicycle Warehouse.’ Change the start time from 6:00pm to 2:00pm, and the duration to 45 minutes long.

Task 3: View a report of Leads by Source.

Task 4: Convert a Lead.

Task 5: Create a new marketing Campaign with the following information:
Campaign Type: Mail
Name: Campaign Times
Status: Planned
End Date: July 31st, 2008
Campaign Owner: John Doe
Budget: 7000

Task 6: Add a new Contact named George Hanson. Add in and/or check the following information:
First Name: George
Last Name: Hanson
Email: georgehanson@google.com
Contact Owner: John Doe
Account Name: Old & New Cycles
Mailing City: St. Louis
Mailing State: Missouri
Description: Met George at the IXDA conference, he is interested in purchasing a new bicycle.

Task 7: Add 2 new tasks with the following information:
Subject of Task 1: File Reports by Friday 5pm
Status: Not Started
Priority: Normal
Assigned To: John Doe
Contact Name: Doris Thurlow
Subject of Task 2: Set up Meeting with Drew ASAP
Status: Pending
Priority: Normal
Assigned To: John Doe
Contact Name: Bill Green

Tasks were then randomized into three randomized task lists to control for participant fatigue, the learning curve, and limited session duration. Participants were then randomly assigned to one of the three task list versions. All twenty-two one hour sessions were conducted by the same two researchers, who either facilitated or took comprehensive notes using Morae Observer. All sessions were also conducted remotely using a web meeting tool and a conference call.

Time on task and number of errors were collected during each task. Following completion of each task, users were directed to the online survey where they selected the Emocard which best represented their initial emotional reaction to the task. They then continued onto the PAD scale, after which participants were asked for their qualitative feedback. This procedure was repeated for each task.

Analysis and Results

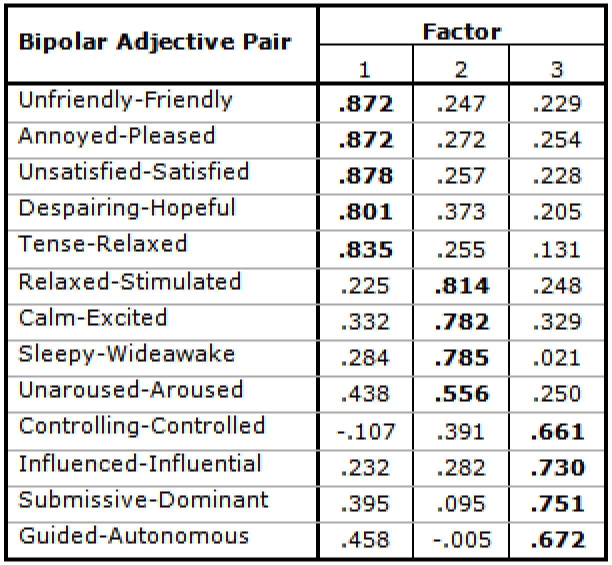
Data was analyzed in two ways for the purposes of this study. First, we analyzed responses across all respondents, regardless of interface, to assess the validity of the PAD scale. Second, we analyzed all data by interface and task to compare responses. Emocard data was analyzed through visual representations of participant selections.

PAD Semantic Differential Scale Analysis

Mehrabian and Russell reported that their PAD scale fell evenly into the emotion dimensions of Pleasure, Arousal, and Dominance [17]. To determine if similar dimensions could be extracted from our study data, a factor analysis of all participant responses was conducted. Factor analysis is a statistical way to look at correlations between items to determine underlying dimensions represented by the variables. Items that are highly correlated are assumed to be measuring similar dimensions. For example, the words car, truck, and minivan may all be grouped together to describe an underlying factor called Automobiles.

As can be seen in table 2, factor analysis resulted in a breakdown of the data into three distinct dimensions, shown by the columns Factor 1, Factor 2, and Factor 3. Adjective pairs that have the highest value in each factor column are grouped together as best representing that dimension. Therefore, looking at the column for Factor 1, the first five adjective pairs with the highest values are grouped together. Factor 2 is best represented by the following four adjective pairs, and Factor 3 by the last four adjective pairs.

The factor analysis breakdown of the original bipolar adjective pairs from our study was identical to those determined by Mehrabian and Russell. Our additional pairs of Tense-Relaxed and Unfriendly-Friendly best represented Factor 1. Factor 1 is therefore equivalent to the Mehrabian and Russell's dimension of Pleasure, Factor 2 is that of Arousal, and Factor 3 may be termed Dominance. These three dimensions were successfully applied to a comparison of interfaces.



Bipolar Adjective Pair	**Factor**		
	1	2	3
Unfriendly-Friendly	**.872**	.247	.229
Annoyed-Pleased	**.872**	.272	.254
Unsatisfied-Satisfied	**.878**	.257	.228
Despairing-Hopeful	**.801**	.373	.205
Tense-Relaxed	**.835**	.255	.131
Relaxed-Stimulated	.225	**.814**	.248
Calm-Excited	.332	**.782**	.329
Sleepy-Wideawake	.284	**.785**	.021
Unaroused-Aroused	.438	**.556**	.250
Controlling-Controlled	-.107	.391	**.661**
Influenced-Influential	.232	.282	**.730**
Submissive-Dominant	.395	.095	**.751**
Guided-Autonomous	.458	-.005	**.672**

table 2. Factor analysis of study data from the PAD Scale validated its measure of the three dimensions of emotion.

A scale reliability analysis was also conducted, which indicated high reliability with a Cronbach's Alpha of .960. Analysis also indicated that the dropping of the items Unaroused-Aroused and Controlling-Controlled would increase overall scale reliability. This finding is relevant for future research applying the PAD scale to interface design.

Analysis of the PAD scale thus validates that Mehrabian and Russell's scale assesses the three dimensions of emotion. It also indicates that scale application may be extended beyond that of advertising and marketing to the domain of software interfaces.

Interface Comparison

An independent sample t-test was used for analysis of the data to compare the two interfaces.

First and foremost, no significant differences were found between interfaces using the usability measures collected in the study (table 3).

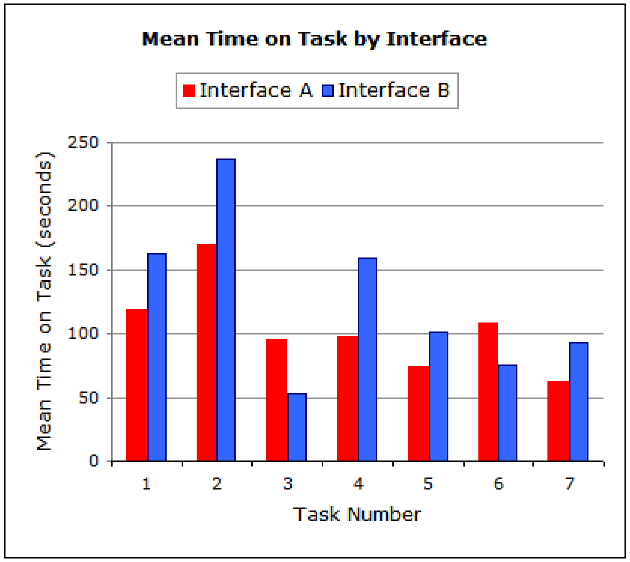


table 3. Although differences in means were observed, statistical analysis indicated that there were no significant differences in usability metrics (time on task shown above) between Interface A and Interface B.

Neither time on task nor number of errors was significantly different between the interfaces when analyzed both overall and by individual task ($p > .05$).

Analysis of the PAD scale, however, did show significant differences in participants' emotional responses between interfaces. Overall, the Interface A was significantly rated by participants as being more Satisfying and Friendly ($p < .05$). When analyzed by task, users rated Interface A as more Pleasing and Relaxing for three out of seven tasks ($p < .05$). Participants therefore found that Interface A elicited a more positive emotional experience than Interface B, even though usability metrics indicated that the two interfaces were nearly identical.

Emocard responses were compared between the two interfaces for each of the seven tasks. Since statistical analysis was not possible, we displayed Emocard data in an innovative visualization (figure 2). As can be seen in the figure, Interface A received an equal amount of Emocard selections in the Excited-Pleasant and Neutral-Pleasant quadrants. In contrast, Interface B received an equal number of Neutral-Unpleasant and Neutral-Pleasant Emocard selections. Clear differences between how users immediately reacted to the interfaces can be identified.

To better understand underlying causes for varying emotional responses, additional information was gathered from session usability notes and respondents' qualitative feedback. Interestingly, use of the PAD scale and the Emocard measures encouraged participants to provide feedback in emotional terms.

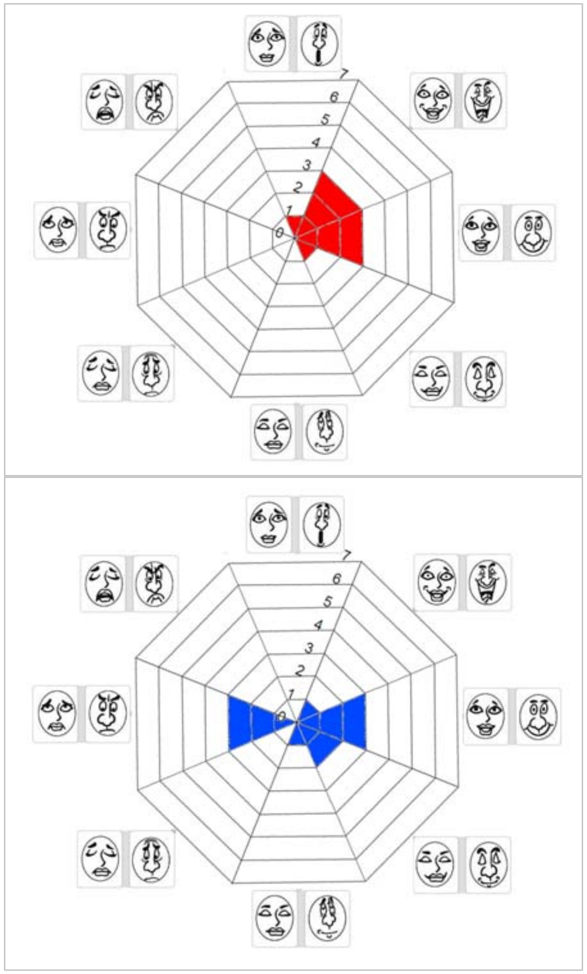


figure 2. Emocard data visualizations for a sample task between Interface A (top) and Interface B (bottom). The visualizations show clear differences in users' immediate emotional responses to the task.

While feedback in usability testing may often include emotional responses, it often depends on the participant and individual differences in how they verbalize feedback. Use of the PAD scale thus seemed to encourage all participants to provide a greater degree of emotional responses than would be observed in a typical usability session. Three sample participant quotes are provided below:

"It took me a while to find the [content]… I chose the slightly perplexed face… after exploring I found the [content] but initially it was a bit frustrating."

"I struggled to find information about the topic, it wasn't intuitive. I searched in several areas… the [interface] should have been linked and said something like 'ok the info isn't here but try this instead."

"I absolutely hate when I see something red that pops up and doesn't tell me anything... It makes me feel stupid. It drives me up the wall. I put a sad face, because it makes me kind of sad… I had a strong negative reaction to that. It was kind of unexpected, [Interface B] had a nice clean interface then this red blinking error popped up out of nowhere. It made me kind of tense."

As can be seen from these few quotes, the qualitative data we collected was both rich in content and often emotionally charged.

Discussion

Positive user experience is often traditionally indicated through usability metrics, even though usability is only one facet of the entire user experience. This strong reliance is encouraged by the numerous methodologies and metrics that have been developed to assess usability, such as number of errors, time on task, heuristic evaluations, usability testing, etc. With these tools, usability is simply the *easiest* component of user experience to assess.

However, researchers miss other key components of user experience by focusing solely on usability. As we have shown in this paper, users also have qualitative emotional experiences when using a product or interface. These emotions may not only be central to how a user judges the overall product experience, but may also affect how a user perceives its usability.

The study described in this paper provided clear evidence of how studying emotion can add value to traditional user experience research. If we had utilized only the usability metrics of time on task and number of errors as measures of user experience -- and believed these measures to be comprehensive indicators of user experience -- we would have concluded that the quality of the user experience for both interfaces were nearly identical. This conclusion, however, would have been incorrect, and, at the very least, incomplete. Differing emotional response to the two interfaces demonstrated that there were significant distinctions between the two interfaces beyond just that of usability.

Since the usability of the two interfaces in our study was nearly identical, we deduced that the significant difference in emotional experience must be due to something *other* than usability.

There is likely a complex and interesting relationship between usability and users' emotional response. As noted earlier, perceptions of usability are often affected by emotion: a happy user may judge a product more usable than an unhappy user, even when using the same product. Additionally, the actual usability of a product will almost certainly have an effect on one’s emotional state. A difficult-to-use product will negatively affect a user’s emotions following attempted usage. Future research into what specific elements of the user experience cause differences in emotional response would be a natural extension our research. Future study could also explore the interactions between actual usability, perceived usability, and emotion.

Emotion is a particularly challenging thing to study. It is a highly subjective, fluid, and completely qualitative construct. In this paper, we learned that a proper study of emotion required a certain amount of “fuzziness” in our research. Early on in our study, it occurred to us that utilizing overly artificial quantitative metrics to define particular emotions would not accurately represent true human emotional experience. Because of this, we were especially attracted to use of the Emocards, since the faces allow for and encourage the fuzziness necessary for realistic emotional measurement.

The purpose of our case study was to demonstrate the value of studying emotion and to test metrics for this purpose. Our scope did not involve research into *how* interfaces might be improved based upon the results of practitioners testing for emotional response. However, we believe that the utilization of these metrics will open up opportunities for HCI practitioners to incorporate fruitful and insightful emotional study into their process. Moreover, interaction designers of software interfaces may best be able to utilize the results of emotional tests and incorporate them into future interface design iterations.

Despite the complexities and issues involved in studying emotional response, it is ultimately the *relevance* of emotion in overall user experience that justifies its inclusion in HCI research.

Acknowledgements

We thank the User Experience team at salesforce.com and Clifford Nass at Stanford University for all their help, support, and interest in this research.

References

[1] Bradley and Lang. Measuring emotion: the Self-Assessment Manikin and the Semantic Differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*, 1 (1994).

[2] Brave, S. and Nass, C. Emotion in human-computer interaction. In J. Jacko & A. Sears (Eds.), *Handbook of human-computer interaction*, Lawrence Erlbaum Associates (2002), 251-271.

[3] Crane, E. A., Shami, N. S., and Peter, C. Let's get emotional: emotion research in human computer interaction. In CHI '07 Extended Abstracts on Human Factors in Computing Systems, ACM (2007), 2101-2104. DOI= http://doi.acm.org/10.1145/1240866.1240958

[4] De Martino, Kumaran, Seymour, and R. J. Dolan. Frames, Biases, and Rational Decision-Making in the Human Brain Science (2006), 313.

[5] Desmet, P.M.A. Measuring emotion. In M. Blythe, A Monk, K. Overbeeke, & P. Wright (eds). *Funology: From Usability to Enjoyment*. Kluwer Academic Press (2003).

[6] Desmet, P.M.A. Emotion through expression; designing mobile telephones with an emotional fit. *Report of Modeling the Evaluation Structure of KANSEI, 3* (2000), 103-110.

[7] Edell, J. A. and Burke, M.C. The Power of Feelings in Understanding Advertising Effects. *Journal of Consumer Research*, *14* (1987), 421- 433.

[8] Ekman, W. and Friesen, P. *Facial Action Coding System (FACS): Manual*. Consulting Psychologists Press, Palo Alto, CA, USA, 1978.

[9] Erevelles, S. The role of affect in marketing. *Journal of Business Research*, 42 (1998), 199–215.

[10] Hirschman, E.C, and Holbrook, M.B. Hedonic Consumption Emerging Concepts, Methods and Propositions. *Journal of Marketing* 46 (1982), 92–101.

[11] Hirt, E.R., Melton, R.J., McDonals, H.E., and Harackiewicz, J.M. Processing goals, task interest, and the mood-performance relationship: A mediational analysis. *Journal of Personality and Social Psychology* (1996).

[12] Holbrook, M.B., & Batra, R. Toward a Standardized Emotional Profile (SEP) Useful in Measuring Responses to the Nonverbal Components of Advertising. In Hecker & Stewert (Eds.), *Nonverbal Communications in Advertising,* D.C. Heath (1987), 95 - 109.

[13] Igbaria, M. An examination of the factors contributing to microcomputer technology acceptance. *Accounting, Management and Information Technologies*, *4,* 4 (1994), 205 - 244.

[14] Kleinginna, P.R., Jr., and Kleinginna, A. M. A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and Emotion*, *5*, 4 (1981), 345-379.

[15] Kotler, P. *Marketing Management*. Prentice-Hall, Inc, Englewood Cliffs, NJ, USA, 1991.

[16] Leavitt, C. A multidimensional set of rating scales for television commercials*. Journal of Applied Psychology*, 54 (1970), 427 - 429.

[17] Mehrabian, A., & Russell, J.A. *An approach to environmental psychology*. M.I.T. Press, Cambridge, MA, USA, 1974.

[18] Picard, R.W. Does HAL cry digital tears? Emotions and computers. In D.G. Stork (Ed.), *Hal's Legacy: 2001's Computer as Dream and Reality,* M.I.T. Press (1997), 279-303.