Building on the Usability Study: Two Explorations on How to Better Understand an Interface

Anshu Agarwal and Madhu Prabaker

salesforce.com, The Landmark @ One Market St. Suite 300, San Francisco, CA 94105
{aagarwal,mprabaker}@salesforce.com

Abstract. In this paper, we describe two separate studies that improved our ability to understand our users' experience of our products at salesforce.com. The first study explored a methodology of combining expert and novice performance data to yield a measure of intuitiveness. The second study created a methodology that combines both verbal and nonverbal emotion scales to better understand the emotional effect our products have on our users. We present both these methods as expansions on the standard usability study and examples of ways to better understand your users within an industry environment.

1 Introduction

Positive user experience is often considered synonymous with good usability. Indeed, empirical usability studies are often used as the sole indicator of overall user experience. However, our work as usability practitioners has suggested that certain components of user experience – such as *intuitiveness* and *emotional response* – may not be sufficiently measured through the usability study.

1.1 Study One: Defining Intuitiveness

The word "intuitive" is a term that has become increasingly common among user experience professionals. It is used in a manner that suggests that it is, at most, a requisite for a good user experience and, at least, a strongly desired characteristic.

A commonsense definition provided by Oxford American Dictionary is "easy to use and understand". Naumann et al. have derived the more formal definition: "A technical system is, in the context of a certain task, intuitively usable while the particular user is able to interact effectively, not-consciously using previous knowledge" [6]. The key component in both definitions seems to be that an intuitive interface enables users to complete tasks efficiently without high cognitive demands.

Although we have metrics for evaluating the efficiency of an interface (e.g. time on task), it is difficult to easily measure a user's cognitive load while using a product. Our research goal in the investigation that follows was create a more usable definition of "intuitiveness". This definition, in addition to making this useful concept more measurable, aims to allow usability professionals the ability to draw more meaningful and complete conclusions about the effectiveness of their interfaces.

J.A. Jacko (Ed.): Human-Computer Interaction, Part I, HCII 2009, LNCS 5610, pp. 385–394, 2009.

1.2 Study Two: Defining Emotional Response

The topic of emotion has recently attracted increased research attention in HCI studies[1]. Numerous authors have proposed that that emotion may play an important role in user performance and user experience. However, very few “real world” case studies have been conducted to study the role of emotion in an HCI context.

It is important to first define the often vague term "emotion." However, coming up with a precise and scientifically respectable definition of the term is notoriously difficult. As one might imagine, there are many definitions of "emotion" in the relevant literature [4]. Nevertheless, there are two generally agreed on aspects of what actually constitutes human emotion [1]. 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 [1].

2 Study One: Measuring Intuitiveness

Although it is often advisable to ensure that designs work equally well with both novice and expert users, not all systems need to be evaluated with this range of expertise. For most “walk up and use” systems, like movie kiosks, or one-time use systems like installation programs, it may not be necessary to test expert users. However, for most consumer and enterprise software systems, the system must allow experienced users to perform their tasks efficiently and novice users to complete tasks effectively without requiring extensive training or practice.

2.1 Measuring Novice Performance

By performing an empirical usability study and measuring the average task completion time across a group of novice users, we can begin to understand how well a particular design performs. Although task completion times allow us to say, “it took a user *x* seconds to complete a task”, they fail to help us understand whether this time is too long or acceptable.

To provide a more comparative understanding, we often visually compare time across all tasks (fig. 1). From this we can say, “it took an average novice user *x* seconds more to complete Task 3 than Task 4”. Although this is a more meaningful statement, it is still difficult to understand how long a task *should* take.

Fig. 1. An example of a visualization of the average task times for novice users on a system

2.2 Measuring Expert Performance

When designing interfaces, we are often concerned with making tasks as efficient as possible. Although we can gauge expert performance by conducting usability studies with expert users and recording task completion time, this is often challenging in practice. It's difficult to find users who are experts in all aspects of an interface – experts in one functional area are often novices in others. Additionally, an expert may not yet exist for a new design. For these reasons, practitioners have utilized human performance modeling methods to create reliable estimates of task performance time for skilled users. A particular model for expert user performance that has proven to produce highly useful and scientifically valid results is Keystroke-Level Modeling (KLM) [2]. When provided with a description of a task being performed, the model applies human performance estimates to produce a predicted task completion time. For the purposes of this paper it is not essential that we understand how exactly KLM is derived, but rather, that KLM is a relatively quick and low cost way to get expert performance task time data.

Plotting these values results in a chart that shows expected task completion times for expert users (Fig. 2). Another way to look at this is that these values represent the *efficiency limit* of a particular design. Using this we can make statements about the minimum task times imposed by the design; for example, "we expect that Task 5 will take at minimum *x* seconds".

Fig. 2. An example of a visualization of the average task times for expert users on a system

2.3 Deriving a Measure of Intuitiveness

Revisiting Our Definition. Earlier we defined intuitive as being able to "interact effectively, not consciously using previous knowledge". We also showed how we could get measurements of novice and expert performance across a set of tasks using a particular design. Mapping these two concepts on each other yields a more measurable definition – an intuitive interface can be thought of as one that "minimizes the difference between expert and novice task performance". When an expert is using the system, they are not consciously thinking about how to use the system, rather, how to solve the task at hand. In this way, the closer the novice user performance resembles expert performance, the more intuitive the interface can be regarded.

Combining Expert and Novice Performance. In the Figure 3, shown below, we've plotted the task completion time for both novice and expert users. The intuitiveness line shows the difference between the *novice user performance* and the *efficiency limit* of the design.

This is a more meaningful metric than the novice or expert measures alone because it enables us to make statements like, "novice users took *x* seconds longer on Task 8 than our design called for." It's important to not underestimate how much more powerful this statement is in driving design changes; it allows us to more explicitly recognize and dissociate the limits imposed by the design (the *efficiency limit*) from the observed performance data.

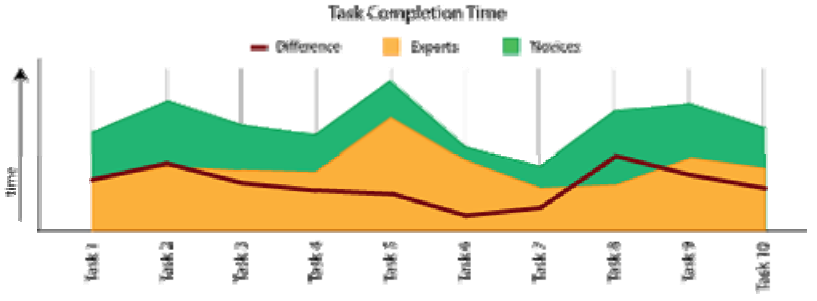

Fig. 3. The expert visualization (fig. 2) has been superimposed on the novice visualization (fig. 1). The difference between the times is shown as a line.

Additionally, because this visualization quantitatively takes into account the inherent difficulty differences between tasks, it enables us to notice phenomenon that's hidden in the novice performance data. For example, although it initially appeared that Tasks 5, 2, 8, and 9 may be problematic because of their relatively long task performance times, Tasks 8, 2, and 9 are the ones that really demand attention – the design actually performed well for Task 5.

While it is quite common in industry to invest the time and money to gather quantitative metrics for novice users, expert efficiency analysis is not always done. By using accurate expert task prediction models, we can achieve deeper insight in our analysis without requiring significant additional resources in our testing phase.

2.4 Empirical Validation of the Method

In order to validate that this method of deriving "Intuitiveness" yielded valuable insight into the performance of a design that was not achieved through the standard usability study, we employed this technique within a comparative study between two versions of a Customer Relationship Management (CRM) application.

To understand the novice user experience, we employed a between-subjects study design where we recruited 18 experienced salespeople to perform a set of 10 common, representative sales tasks (e.g. adding tasks, converting a lead, sharing an opportunity,

etc) as quickly as possibly without committing any errors on one of two CRM applications (Application A and Application B). All participants reported familiarity with each of the sales tasks, but none of them had prior experience with the application they were assigned to.

For each session, participants were presented with the tasks in a randomized order and among the dependent metrics collected were Time on Task and Number of Assists. Because we were focused on understanding a more natural assessment of the time it took novice users to complete a task, we chose to provide assists instead of capturing the number of errors committed[1]. This methodology ensured that all participants completed each task and that our Time on Task metric captured the inherent difficulty novice users had.

To understand the expert performance times for each of the 10 tasks, we performed KLM analysis using the software application, CogTool [7].

Table 1. The table below shows the 10 task times for Novice Users (Empirical), Expert Users (KLM), and the difference between these two (Intuitiveness). The design that performed better has been bolded for each task.

	Novice Performance		**Expert Performance**		**Difference (Intuitiveness)**	
	A	B	A	B	A	B
1. Complete a Task	153.91	**91.22**	**19.82**	21.74	134.09	**69.48**
2. Add a few tasks	**187.75**	197.19	64.81	**55.46**	**122.94**	141.73
3. Edit a contact	**118.06**	118.71	**19.47**	21.15	98.59	**97.57**
4. Convert a lead	**114.26**	184.51	**17.26**	23.75	**97.00**	160.77
5. View reports on leads	**82.93**	105.57	**6.93**	8.82	**76.00**	96.75
6. Share an opportunity	**150.35**	152.65	25.63	**25.27**	**124.73**	127.38
7. Manipulate a calendar entry	231.57	**203.36**	**30.1**	33.65	201.46	**169.71**
8. Manipulate a forecast	**77.68**	93.05	6.82	**6.48**	**70.87**	86.58
9. Create a campaign with leads	292.28	**237.77**	**27.16**	54.72	265.12	**183.05**
10. Search using help	**58.215**	86.33	13.65	**9.07**	**44.56**	77.25

Analysis and Results. Although no statistical difference was found across the overall task performance of novice users, users were statistically faster on Task 1 using Application B ($p = 0.008$)[2]. Application A had a lower expert performance time for six out of ten tasks[3].

[1] An assist was provided when the participant ceased making progress towards the completion of the task. The assist was given such that it only provided the user with enough direction to make it to the next step in the task and only when it became clear that the user was unable to advance to the next step.

[2] We performed a Two-Sample T-Test on the novice, empirical, performance data.

[3] Since the KLM values are not empirically derived, we can consider any difference between the designs as significant.

The value of this method can be seen based on how the conclusions might differ based on the data at hand. Armed with only the traditional, empirical usability study data we might be able to derive that both applications performed equally well with novice participants, though Application B had a more efficient interface for Task 1. Therefore if we are redesigning Application A, we should focus our effort on improving our design for Task 1; since the other nine tasks performed statistically similarly it's unclear as to whether the designs on both are equally good or equally poor. However, once we add the expert performance metric and derive the Intuitiveness metric, we start to see a more interesting and insightful picture. Application B's faster time for Task 1 cannot be attributed to an overall more efficient design – in fact Application A's design allowed for expert users to complete the task faster than Application B's design. Therefore, for Task 1, although Application A was more efficient than Application B, it was less intuitive. In this way we've changed the focus of our redesign efforts from a focus on efficiency to a focus on making it easier for the novice user to accomplish. With this insight, if our task is to redesign Application A we cannot help but notice that, in additional to Task 1, Tasks 7 and 9 should be the focus of our efforts.

3 Study Two: Measuring Emotional Response

Emotion is an inherently complex construct to study. As such, researchers have created many different emotion measurement tools, including verbal measurement tools, nonverbal measurement tools, and physiological measurement tools in an effort to meet this challenge. In this study, our research challenge was to develop an emotion measure that would be quick to utilize, easy to understand, deployable remotely, and easy to incorporate into an empirical usability study.

Given the nature of emotion, it would seem that "fuzzy" nonverbal measures would be most apt to assess emotion. However, most of the nonverbal measures in the HCI literature are either impractical in a "real world" setting, or of unknown validity. We therefore decided to combine an extensively used and validated *verbal* scale with a more experimental *non-verbal* emotion measure to improve the strength of our methodology.

3.1 Verbal and Non-verbal Emotion Measurement

For the verbal component, we chose to utilize the PAD (Pleasure, Arousal, and Dominance) Semantic Differential Scale developed by Mehrabian and Russell [5]. By rating a set of bipolar adjective pairs along a nine-point range, this 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. *Arousal* refers to an emotional state from sleepy to very excited. The final dimension, *Dominance*, refers to the extent to which a person feels unrestricted or free from outside control. We reviewed Mehrabian and Russell's original adjective sets to ensure that the pairs were relevant to interface emotional responses (Table 2).

Table 2. Although we maintained most of the original adjective word pairings of the PAD scale, we revised some pairings to ensure that the scale was concise and relevant to software interface assessment

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		

We selected the Emocard tool by Desmet for the non-verbal component of our measure (fig. 4) [3]. The Emocard tool consists of sixteen cartoon-like faces, half male and half female, in which each face represents a combination of Pleasure and Arousal. We interpreted results in the Calm-Pleasant and Excited-Pleasant quadrants as positive feedback.

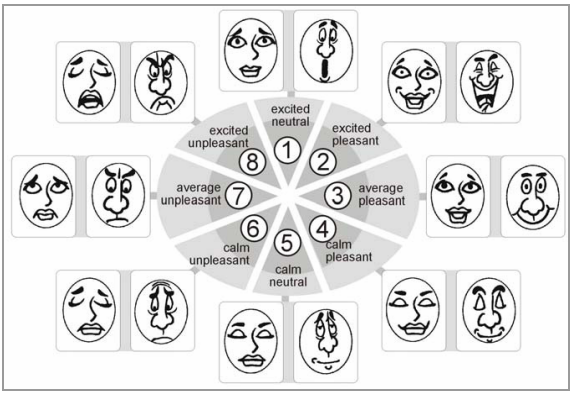

Fig. 4. The Emocard tool was an effective nonverbal measurement of emotional response which used human-like representations of emotion

3.2 Empirical Validation of the Method

In order to validate this methodology, we performed a comparative study between two versions of a CRM application interface. We collected traditional usability measures (*time on task* and *number of errors*), as well as the new dual *emotion measure* we constructed. This measure utilized both the non-verbal Emocards and the verbal PAD scale methods in a linear fashion.

Twenty-two participants, thirteen male and nine female, were assigned to assess one of the two versions of the interface. Although participants had experience with CRM, they had no prior experience with the interface they were evaluating. Seven comparable CRM tasks between the two interfaces were created (e.g. manipulate a calendar entry, view a report of leads by source, create a new marketing campaign, etc). These tasks were representative of typical sales users of CRM interfaces. Tasks were randomized and participants were assigned to one of the three task list versions.

As collected in a usability study, the traditional measures of *time on task* and *number of errors* were collected during each task. This was followed by an online survey where participants selected the Emocard that best represented their initial emotional reaction to each task. Participants then continued onto the PAD scale, and were asked for their qualitative feedback. This procedure was repeated for each task.

Analysis and Results. No significant differences were found between interfaces using the usability measures collected in the study[4]. Neither *time on task* nor *number of errors* was significantly different between the interfaces when analyzed both overall across all tasks 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 user's performance levels in the usability studies were almost identical.

Emocard responses were then compared between the two interfaces for each of the seven tasks (Fig. 5). As can be seen in the figure, clear differences and patterns in how users immediately reacted to the interfaces can be identified. Interface A elicited a more consistently positive response compared to Interface B, which included the selection of a few Emocards that represented negative emotions.

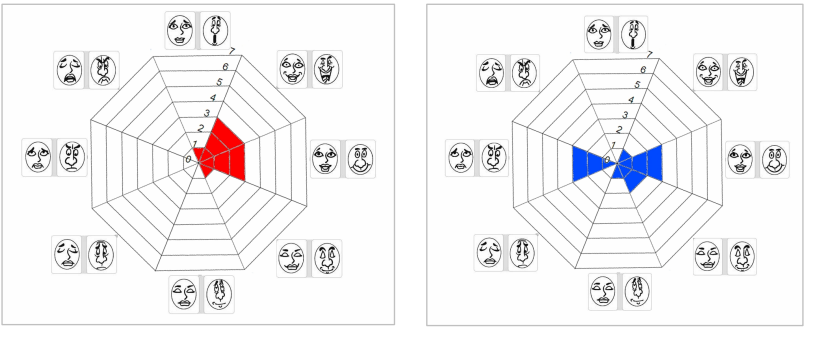

Fig. 5. Emocard selections for a sample task between Interface A (left) and Interface B (right) show clear differences in users' immediate emotional responses

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

Qualitative feedback was also collected for each task. Two 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 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 indicated in these quotes, the qualitative data we collected was both rich in content and often emotionally charged.

3.3 Studying Emotional Response: Considerations

Practitioners might assume that positive emotional response may be adequately indicated through usability metrics. However, the results of this study suggest that this may not be the case. 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. Additionally, 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 goal of this study was to demonstrate the value of studying emotion and to test metrics for this purpose. Utilization of these metrics may help 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 emotion studies to help enhance their interface designs.

4 Conclusion

The two studies outlined in this paper demonstrate how studying emotion and measuring intuitiveness can add value to traditional user experience research. Both studies utilize new methods that practitioners can use to build upon the traditional usability study. Both explorations also yielded significant insight into our understanding of our users' experience with marginal additional effort.

The research efforts discussed here were only initial exploratory studies that merit further research. The intuitiveness measure still demands more empirical testing to validate its ongoing value and accuracy. Although emotional response has been proven a valuable aspect to study, further exploration of how interfaces might be improved based upon the results should be conducted. In the end, these methodologies hope to benefit the user experience community by encouraging practitioners to extend their everyday usability research in search of greater insights.

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

References

1. Brave, S., Nass, C.: Emotion in human-computer interaction. In: Jacko, J., Sears, A. (eds.) Handbook of human-computer interaction, pp. 251–271. Lawrence Erlbaum Associates, Mahwah (2002)
2. Card, S.K., Moran, T.P., Newell, A.: The keystroke-level model for user performance time with interactive systems. Communications of the ACM 23(7), 396–410 (1980)
3. Desmet, P.M.A.: Emotion through expression; designing mobile telephones with an emotional fit. Report of Modeling the Evaluation Structure of KANSEI 3, 103–110 (2000)
4. Kleinginna Jr., P.R., Kleinginna, A.M.: A categorized list of emotion definitions, with suggestions for a consensual definition. Motivation and Emotion 5(4), 345–379 (1981)
5. Mehrabian, A., Russell, J.A.: An approach to environmental psychology. MIT Press, Cambridge (1974)
6. Naumann, A., Hurtienne, J., Israel, J.H., Mohs, C., Kindsmüller, M.C., Meyer, H.A., Husslein, S.: Intuitive Use of User Interfaces: Defining a Vague Concept. In: Harris, D. (ed.) Engineering Psychology and Cognitive Ergonomics, HCII 2007, vol. 13, pp. 128–136. Springer, Heidelberg (2007)
7. The CogTool Project: Tools for Cognitive Performance Modeling for Interactive Devices. Carnegie Mellon University (April 16, 2006), `http://www.cs.cmu.edu/~bej/cogtool/index.html`