AffectButton: Towards a Standard for Dynamic Affective User Feedback

Joost Broekens and Willem-Paul Brinkman
Man-Machine Interaction group, Delft University of Technology
Delft, The Netherlands
joost.broekens@gmail.com, w.p.brinkman@tudelft.nl

Abstract

Emotions are an important aspect of life. Persons have emotions while using products and technology. It is becoming more and more important to be able to assess these emotions for multiple reasons: (a) to develop better products, (b) to better understand how the user interacts with products, and (c) because the affective state of the user is of importance to the product itself (e.g., in the case of social software, persuasive computing, recommendation). In general there are two ways of extracting affective user feedback: explicit and implicit. Here we present a new interface component that enables users to give explicit affective feedback in a flexible and dynamic way. We call this component the AffectButton. Based on statistical analysis of affective user input gathered with the AffectButton in three user studies, we present evidence that users can use the button effectively to enter affective feedback. Furthermore, the feedback is reliable and valid.

1. Introduction

Emotions play an important role in our lives. Emotion and affect influence how we think, what we like as well as how we react to and reason about events As such, emotions also play a role in how we use products and technology. In the recent years the measurement of experience, of which emotion is an important part, has become very important [17]. A key aspect of this research is the development of valid measurement tools for real-life experience measurement. Often, this new way of product design research is called a *Living Lab* approach. The main aim of such an approach is to better design products for people by capturing how they use the product in their daily life. This is a markedly different approach from the traditional controlled laboratory experiment. As emotion is a key aspect of experience, it is important to capture a user's emotion in a valid, reliable and easy to use way. This paper addresses this question.

A second recent development is the emergence of social software. Such products explicitly mediate or use human-to-human communication. Network sites such as *Facebook* facilitate human contact. Blogs facilitate topic-based communication and discussion, and wiki's facilitate collaborative creation. As emotion is important in human-to-human communication, measuring and presenting emotion is an important field of study for this emerging field [4].

Third, there are applications that could use human affect to function. For example, content sharing systems such as *Tribler* [23] facilitate human-to-human content recommendations in a social way. Recommendations are made based on collaborative filtering, a user-to-user content matching mechanism based on overlapping interests in content items. In a sense such mechanisms try to capture the experience and emotion a user feels when consuming a content item. Other systems that use affective feedback include e-learning and pervasive computing systems [9] as well as virtual reality therapies [11], training simulators and tutoring systems [7]. Such systems would benefit from a valid, easy-to-use and reliable emotion measurement tool.

There is a long history of emotion measurement in psychology. Typical instruments used are questionnaires, adjectives (e.g., [25]), images [14], role playing, as well as a set of tools that measure different physiological modalities such as heart rate and skin conductance. Further, emotion recognition has been a topic of research for quite some time in the area of affective computing (See ,e.g., [20, 21, 22]). On the one hand the measurement tools are well-validated and readily available. On the other hand, many of these tools are not useable in real-life situations and many have only been tested in experimental laboratory settings. A recent review and evaluation of 5 different emotion measurement tools concluded that many are essentially not suited for measuring emotions in the mobile computing domain [10].

Affective user feedback can be divided in two categories: explicit and implicit. Implicit methods sense the behavior of the user (face, body, heart rate, mouse-movement, etc.) and deduce an emotion. Explicit methods ask the user to input affective feedback directly. This paper presents a new digital measurement tool for explicit affective feedback: the *AffectButton*. The tool is an interface component that functions, looks and behaves like a button but enables a user to input a dynamic (i.e., graded intensity and mixed) affective state. The component is easily integrated in an interface and, as our results show, is understood by most users as well as results in valid and reliable affective input. We now motivate why such a component is needed.

978-1-4244-4799-2/09/$25.00 ©2009 IEEE

Many approaches exist towards explicit emotion feedback by means of digital systems [5, 10, 27]. However, these approaches typically have a fundamental tradeoff between precision and measurement speed/ease of use [10]. This means that the more detailed the feedback, the more effort involved for the user and therefore the less likely users will adopt the method as a common way of entering affective feedback. In our approach we specifically aim for both precision and speed/ease of use. To evaluate ease of use, we measure the amount of effort involved in using the AffectButton.

Further, many methods ask a user to input categorical emotions with or without intensity [5, 10, 27]. The benefit is that discrete emotions are easy to interpret by the user giving the feedback and the person or system interpreting the feedback. The drawback is that mixed emotions are difficult to express as there is no logical "emotional continuum" between categories.

Other approaches are able to, in principle, extract detailed affective information in a non-invasive manner but involve the use of human observers to evaluate the feedback and are more focused on measuring human emotion during the process of product development [12, 13]. Such approaches give detailed affective feedback but are not suitable for online affect measurement: i.e., affect measurement aimed at getting real-time affective feedback in a format that is usable by a computer system. Our goal is the latter. We aim for a simple, easy to use device that can be used for affective feedback without the need for human observer intervention.

Several methods exist that are based on the Self-Assessment Manikins (SAM) [2]. Key in these methods is that they measure emotion factors (pleasure, arousal and dominance) directly and separately. For each factor the user selects a picture from a set of pictures showing emotional faces that express different intensities for that factor. Although the SAM method is by now well-validated, a potential unresolved limitation is that users must understand the three emotion dimensions before they can use the method. A second drawback is that the method takes up a considerable amount of screen space. The AffectButton addresses both issues.

Finally, not many approaches emphasize the fact that affect measurement must be reliable (i.e., is my measure precise) and valid (i.e., do I measure what I want), and preferably generic (usable in different interfaces).

To summarize, the motivation for developing the AffectButton is that there currently is no measurement method for affect that is thoroughly evaluated with respect to validity and reliability, simple to use and understand for users, easy to embed, of which the data is machine interpretable, and is able to measure mixed and graded intensity emotions. In this paper we discuss three main things: the validity and reliability of the AffectButton as an affect measurement device, and usability of the button in context. We describe three experiments that were conducted to study this. The results show that the AffectButton can measure affective feedback quickly, that users understand the button and like the concept, that affective feedback by means of the AffectButton is valid and reliable and that the button can be used in context. Finally we discuss the benefits and limitations of our approach.

2. AffectButton: Pleasure-Arousal-Dominance-Based Feedback

Emotion is a complex topic, and agreement on one solid definition does not really exist. We do not detail the topic of emotion, as many excellent works have been published from different perspectives (see, e.g., [6, 15, 18, 19, 21, 24, 28]. We explain how to interpret what the AffectButton measures in relation to emotion, and above mentioned references can be seen as "collective source".

Typically, affect refers to the underlying core of emotion, mood and affective attitude towards persons and things. Emotion, mood and affective attitude are different but strongly related and influence each other. In general, emotion is related to facial expression, feeling, cognitive processing, physiological change and action readiness. Furthermore, emotion refers to a short but intense episode that, in addition to the previously mentioned aspects such as facial expressions, is characterized by "attributed affect to a causal factor". An emotion is a noticeable if not powerful experience. For example, I feel (and notice I am) happy about seeing an old friend. In contrast, mood refers to a silent presence of moderate levels of affect. I can feel frustrated for half a day without knowing why. Mood is not (consciously) attributed to a causal factor. Affective attitude refers to how one generally feels about something or someone, not specifically because of that thing or person. For example, *I like popular science books*. To complicate matters a little, affect is also used as commonplace term for everything that has to do with the above.

There are several theoretical views on how to think about emotion. The following categorization that uses two axes is particularly useful. The first axis defines the level of abstraction at which emotion is studied: social, psychological, biological, and physiological. The second axis defines the way emotion is represented: categories of emotion, components that form an emotion, and principle factors. For example, the well-known six basic emotions as proposed by Paul Ekman are categorical (fear, anger, happiness, etc.). Cognitive appraisal theories are componential, as these describe emotion as a combination of the activation of different sub processes (evaluation of an event in terms of novelty, goal conduciveness, etc.). On the other hand, Russell [26] proposes a description of emotion using two continuous factors (Pleasure, Arousal).

Disregarding these different views, many emotion researchers agree upon two common affective factors that are useful to describe a mood, emotion or attitude: *valence* and *arousal*. The difference in opinion is not so much about these factors but about how to interpret what they are. Are these factors the emotion, do they represent something real in the brain and if so which brain areas are involved, are they independent (orthogonal), are they artifacts of statistical analysis of many factors, etc.

One of the goals of our research is to have a measurement device to input mixed and graded intensity affective feedback, regardless of to what that feedback relates (attitude, emotion, and mood). We have chosen for a dimensional approach that relates to core affect as explained above. Because a substantial number of emotions cannot be represented clearly as points in this 2D affective space, we have used a related theory as basis for the AffectButton. This theory uses three factors, Pleasure (i.e., valence), Arousal and Dominance [16] (PAD), and is more expressive a model for affect but less generic as it is unclear if the Dominance dimension is a fundamental element of core affect. The PAD factor-based theory states that every object/emotion/mood/etc has a mapping to a PAD value triple. What is not the case is that every PAD triple has a unique emotion attached to it, i.e., the reverse mapping. So, the mapping is many (things/mood/emotion) to 1 (PAD triplet).

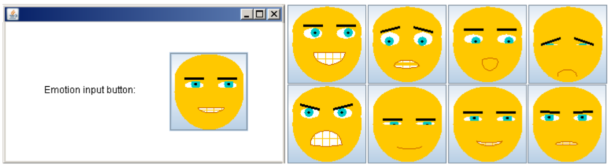

Figure 1. The AffectButton and several examples of affective states: happy (PAD=1,1,1), afraid (-1,1,-1), surprised (1,1,-1), sad (-1,-1,-1), angry (-1,1,1), relaxed (1,-1,-1), content (1,-1,1), frustrated (-1,-1,1). Note that labels are exemplary.

The AffectButton was developed as a simple button (implemented in Java) and is available from http://www.joostbroekens.com. It is scalable in size and is a square (Fig. 1). It does not unfold or pop-out, and can therefore be considered a static element in an interface. The button itself renders a face that changes directly according to the mouse position in the button as well as the scroll wheel. The mouse x and y within the button and the scroll wheel define the values on the dimensions Pleasure, Dominance and Arousal respectively. Arousal is often equated with activeness (alertness), and as such moving back and forth with your body (low alertness versus high alertness respectively) relates to the mouse wheel rolling back and forth. All values can be between -1 and 1. The user can therefore select any affective value from the PAD space using the mouse within the button. Affective values are represented by the rendered facial expressions, so the user selects an emotional expression by clicking the button. The button output is an event with the P, A and D values the user clicked as well as a typical, but by no means unique, emotion label that would go with these coordinates in PAD space. The face is rendered using a mechanism that is comparable to the one used in Kismet [3]. Nine prototypical expressions are defined, one for each extreme in the PAD space (e.g., 1, 1, 1 is Happy) and one for the center, neutral emotion. The expressions are defined in terms of eyebrow, eye and mouth configuration of the face. Five of the 8 extremes in the PAD space map to Ekman's basic emotions and as such we have taken these as prototypical expressions. The remaining three prototypical expressions have been chosen to match as closely as possible the emotion words defined in the three remaining PAD space extremes [16]. Based on the PAD coordinates, the face displayed is interpolated between these nine prototypical expressions. Therefore, a user can enter mixed emotions (e.g., confused) as well as low and high intensity prototypical ones (e.g., little happy, elated).

To evaluate a bare-minimum button that can still reliably measure affective quality, the face is rudimentary and gender neutral. This is in line with for example the design of the iCat robot [1] containing the same number of facial features. We take this approach for two main reasons. First, generic applicability of the button enforces us to not “fancy up” the button; the design should not distract from the button's purpose. Second, from a research point of view it is better to start simple and add more features as needed. Such an approach gives a clearer view on what the different features may add to the measurement capabilities.

3. Evaluation Experiments: Validity, Reliability, Contextual Use

A key concern in the development of any measurement instrument is its validity. Is the instrument measuring what it should? Doubts about this can create serious reservations against using a measure. A good example of such concern has been the discussion about the validation studies of some common usability evaluation methods [8]. With this in mind we have evaluated the AffectButton in three different experiments. First we evaluate the usability of the button in context, where the pleasure and dominance axes are controlled by the mouse x and y coordinates in the button, and the mouse wheel controls the arousal axis. In this experiment we evaluate how users perceive the button when affectively scoring holidays. In the second experiment we evaluate the reliability and validity of the button. In the third experiment we evaluate the validity and reliability of a simplified version where the mouse wheel is not needed. In this version, the 3D PAD space is flattened to a 2D space, with Arousal control at the extremities of the 2D Pleasure Dominance plane.

3.1. Experiment 1: 3D AffectButton Used in Context

In this experiment we study the usability of the AffectButton. We wanted to find out if users like the button, how much effort it costs to use it and if it produces useful data. We asked users to affectively score preferences. Preferences, in our case preferences for holidays, can be considered affective constructs, as they are about liking versus disliking objects and liking is a fundamental affective quality (see e.g., [16, 26]). In this experiment we conducted a study with one experimental variable (2 conditions). Users scored holidays one by one using a 9-points scalar (baseline condition) and using the 3D AffectButton. We wanted to study the following questions. First, do users like to give affective feedback. Second, how much perceived effort is involved, and third, is affective feedback useful to predict the “ideal” item ordering of a user (each user was asked to completely sort all 27 available holidays). This last aspect was evaluated to assess the usefulness of the data produced by the AffectButton in context.

The setup was as follows. The material consisted of a set of 27 cards showing holidays. Furthermore, we used a computer interface that included 2 different tasks. In these tasks participants were asked to rate 9 holidays, randomly generated out of the previously mentioned set of 27 holidays, one at a time. Rating was done using either a 9-point Likert scale from like to dislike or with the AffectButton. We tested 32 Dutch participants (one had used the AffectButton before), 10 female and 22 male, who were mainly students and researchers within the field of computer science aged between 21 and 31. Each participant had to do all three tasks (scalar rating, affective rating and sorting the set of 27 cards from most preferred to least preferred). The order of the three tasks was counterbalanced per participant. After each task the user had to indicate the perceived level of effort and liking of the task on a 7-point Likert scale.

3.2. Experiment 1: Results and Discussion

The first step of the analysis focused on the effect of the measurement tool on effort and on how much people like using the tool. A multivariate analysis of variance (MANOVA) with repeated measures was conducted with as within-subject variable the type of measure instrument (AffectButton or Likert scale) and as dependent measure the effort rating and the like rating. The analysis found a significant main effect (F(2,30)=24.00; p. < 0.001) for measure type. Separate univariate analyses on the effort rating also revealed a significant (F(1,31)=46.32; p.<0.001) main effect for measure type, but not in the linking rating. This means that affective feedback using the AffectButton is associated with a higher perceived effort in preference elicitation (affective effort=3.9, sd=1.6; scalar effort=2.8, sd=1.2), but that the amount of effort involved did not influence the likeability of the button, at least in our particular case (affective liking=4.2, sd=1.6; scalar liking=3.9, sd=1.3). Finally, we wanted to evaluate if affective feedback adds useful information compared to scalar feedback in order to predict a user’s

Mean affect word	3D Button mean (std)			2D Button mean (std)			Mehrabian mean (std)		
	P	A	D	P	A	D	P	A	D
suspicious	-.27(.36)	-.78(.57)	.23(.46)	-.06(.59)	-.76(.45)	.13(.47)	-.25(.23)	.42(.21)	.11(.32)
alert	.14(.22)	.44(.81)	.06(.50)	.11(.40)	-.40(.51)	.48(.53)	.49(.25)	.57(.20)	.45(.26)
fearful	-.72(.16)	.92(.21)	-.70(.16)	-.76(.20)	.29(.32)	-.73(.23)	-.64(.20)	.60(.32)	-.43(.30)
distressed	-.59(.28)	-.50(.63)	-.51(.28)	-.62(.16)	-.57(.31)	-.47(.26)	-.61(.17)	.28(.46)	-.36(.21)
protected	.52(.50)	-.44(.80)	.16(.45)	.34(.33)	-.86(.48)	.02(.37)	.60(.35)	-.22(.37)	-.42(.40)
happy	.71(.20)	.31(.87)	.76(.18)	.77(.15)	.11(.60)	.75(.33)	.81(.21)	.51(.26)	.46(.38)
angry	-.71(.28)	.33(.86)	.76(.14)	-.87(.06)	.39(.43)	.80(.18)	-.51(.20)	.59(.33)	.25(.39)
masterful	.44(.31)	.22(.82)	.54(.33)	.29(.36)	-.25(.66)	.67(.27)	.58(.25)	.44(.27)	.69(.25)
grateful	.69(.19)	-.36(.72)	.13(.53)	.66(.21)	-.49(.55)	.41(.32)	.64(.23)	.16(.22)	-.21(.34)
terrified	-.78(.20)	.97(.10)	-.82(.12)	-.81(.23)	.66(.13)	-.94(.02)	-.62(.20)	.82(.25)	-.43(.34)
fascinated	.65(.25)	.67(.57)	-.46(.38)	.66(.32)	.00(.62)	-.23(.60)	.55(.22)	.51(.23)	-.07(.35)
frustrated	-.67(.30)	.28(.85)	.43(.35)	-.65(.19)	-.43(.57)	.20(.41)	-.64(.18)	.52(.37)	-.35(.30)
interested	.37(.28)	.44(.72)	.00(.39)	.53(.24)	-.24(.48)	.19(.68)	.64(.20)	.51(.21)	.17(.40)
irritated	-.48(.30)	-.42(.81)	.38(.32)	-.59(.28)	-.55(.61)	.37(.23)	-.58(.16)	.40(.37)	.01(.40)
jealous	-.50(.38)	-.03(.96)	.18(.40)	-.45(.29)	-.83(.32)	.20(.32)	-.32(.32)	-.11(.27)	.05(.29)
powerful	.21(.44)	.58(.65)	.57(.36)	.33(.33)	.31(.60)	.83(.18)	.54(.26)	.45(.36)	.73(.25)
curious	.53(.29)	.58(.62)	-.02(.46)	.58(.25)	-.16(.58)	.18(.61)	.22(.30)	.62(.20)	-.01(.34)
impressed	.65(.24)	.89(.26)	-.53(.28)	.64(.32)	.01(.68)	-.54(.44)	.41(.26)	.30(.25)	-.32(.34)
uninterested	-.10(.38)	-.81(.39)	-.32(.29)	.02(.29)	-.95(.16)	-.23(.33)	-.47(.26)	-.50(.22)	-.08(.24)
discontented	-.41(.30)	-.72(.58)	.14(.49)	-.34(.25)	-.89(.27)	-.03(.42)	-.53(.19)	-.16(.41)	-.26(.30)
excited	.73(.18)	1.00(.00)	.07(.72)	.79(.24)	.61(.16)	.29(.76)	.62(.25)	.75(.20)	.38(.29)
relaxed	.59(.22)	-.36(.80)	.12(.43)	.45(.25)	-.87(.27)	.10(.38)	.68(.30)	-.46(.38)	.06(.49)
guilty	-.45(.32)	-.33(.80)	-.40(.23)	-.29(.37)	-.80(.52)	-.36(.29)	-.57(.19)	.28(.38)	-.34(.28)
serious	-.03(.11)	-.17(.88)	.05(.27)	-.03(.21)	-1.00(.00)	.05(.36)	.27(.22)	.24(.22)	.42(.27)
triumphant	.59(.28)	.69(.63)	.69(.22)	.67(.21)	.02(.57)	.79(.14)	.69(.23)	.57(.19)	.63(.26)
astonished	.20(.45)	.94(.19)	-.57(.51)	.34(.59)	.48(.28)	-.89(.07)	.16(.26)	.88(.19)	-.15(.26)
sad	-.58(.26)	-.89(.16)	-.66(.25)	-.46(.12)	-.96(.13)	-.52(.20)	-.63(.23)	-.27(.34)	-.33(.22)
humiliated	-.54(.25)	-.47(.61)	-.40(.40)	-.50(.27)	-.51(.42)	-.55(.21)	-.63(.18)	.43(.34)	-.38(.30)
defeated	-.59(.26)	-.53(.66)	-.58(.28)	-.54(.25)	-.60(.35)	-.62(.17)	-.61(.24)	.06(.39)	-.32(.23)
bored	-.05(.44)	-.72(.42)	-.41(.34)	-.02(.08)	-.95(.13)	-.40(.23)	-.65(.19)	-.62(.24)	-.33(.21)
confused	-.27(.26)	.36(.87)	-.34(.41)	-.44(.39)	-.15(.70)	-.30(.51)	-.53(.20)	.27(.29)	-.32(.28)
hostile	-.58(.33)	.00(.99)	.76(.13)	-.75(.16)	.20(.40)	.81(.18)	-.42(.31)	.53(.36)	.30(.32)

Figure 2. Average PAD scores from experiment 2 and 3. Button 3D is the wheel-mouse version, 2D is without wheel mouse, Mehrabian are known scores as published in [16].

Affect Correl.		3D Button (significance)			2D Button (significance)			Mehrabian (significance)		
		P	A	D	P	A	D	P	A	D
3D Button	P	1.00(-)	.31(.08)	.24(.19)	.98(.00)	.14(.44)	.34(.06)	.93(.00)	.03(.86)	.43(.01)
	A	.31(.08)	1.00(-)	.02(.90)	.29(.11)	.87(.00)	.07(.72)	.40(.02)	.76(.00)	.28(.12)
	D	.24(.19)	.02(.90)	1.00(-)	.19(.31)	.10(.60)	.95(.00)	.38(.03)	.14(.43)	.73(.00)
2D Button	P	.98(.00)	.29(.11)	.19(.31)	1.00(-)	.11(.53)	.28(.11)	.90(.00)	.04(.83)	.42(.02)
	A	.14(.44)	.87(.00)	.10(.60)	.11(.53)	1.00(-)	.11(.54)	.22(.22)	.80(.00)	.27(.14)
	D	.34(.06)	.07(.72)	.95(.00)	.28(.11)	.11(.54)	1.00(-)	.48(.01)	.15(.42)	.81(.00)
Meh-rabian	P	.93(.00)	.40(.02)	.38(.03)	.90(.00)	.22(.22)	.48(.01)	1.00(-)	.18(.32)	.60(.00)
	A	.03(.86)	.76(.00)	.14(.43)	.04(.83)	.80(.00)	.15(.42)	.18(.32)	1.00(-)	.30(.10)
	D	.43(.01)	.28(.12)	.73(.00)	.42(.02)	.27(.14)	.81(.00)	.60(.00)	.30(.10)	1.00(-)

Figure 3. Correlations obtained from experiment 1 and 3. Button 3D is the wheel-mouse version, 2D is without wheel mouse, Mehrabian are known scores as published in [16].

holiday sorting. We conducted a backward stepwise regression analysis with holiday Likert rating and pleasure, arousal, dominance ratings as independent variables over all items to predict the item ranking using the user's baseline preferences given by the 27-holiday card ordering task. The regression analysis with holiday-ranking as dependent variable resulted in a significant model (F(2,285)=110; p.<0.001) with a correlation between actual ranking and predicted ranking of r=0.66. The model included as significant parameters the item Likert rating (Beta=-0.55; t=-9; p.<0.001) and the item pleasure rating (Beta=-0.15; t=-2.5; p.=0.012). This means that, even in a simple linear model, affective feedback (in our case *pleasure*) adds something unique in order to predict user preferences and can therefore be used to better understand human preferences.

Concluding we can say two things. First, users understood the affective feedback mechanism as it added useful data to predict holiday rankings. Second, although users perceive it to be more effortful, they do not seem to be bothered by the effort involved, at least while the button still has novelty value. We do not know if this effect remains when the novelty value wears off.

3.3. Experiment 2: 3D AffectButton with Mouse Wheel Control

In this second experiment we evaluated the 3D AffectButton with respect to concurrent validity (can the AffectButton replicate PAD values of emotion words that have previously validated PAD values), internal reliability (do different users score an emotion word approximately the same), and usability (we measured the time needed to select an emotion using the button).

To evaluate these issues we invited 50 users to participate in an experiment they could do at a time they preferred at their own PC. Of this selection, 16 users responded and participated in the experiment.. Users were presented with a short written introduction to the experiment and one sentence on how to use the button (i.e., use the mouse AND the mouse wheel). The goal in this task was to find the best matching expression for a word using the AffectButton. Mehrabian [16] presents a long list of validated emotion words positioned along the three axes. These words have associated average PAD values obtained by psychological experimentation. This enables us to analyze the concurrent validity of our input mechanism using an already validated set of affectively scored words. As we know these "normal" PAD values of each word, presenting the word and asking a user to select a matching expression is a benchmark test that can be used to analyze the buttons concurrent validity. Users had to match 32 emotion words presented in English and Dutch (translated according to standard dictionary), one by one and in random order. These 32 emotion words were carefully chosen from a longer list of about 150 words [16] according to the following criterion. The 32 words are words with the least amount of variation in the PAD means (i.e., the words that are most consistently perceived in terms of PAD values) and the words fill the complete 3D PAD space. These 32 words thus function as benchmark stimuli used to validate the input of the user. If the button is a valid affect input device, the user's input via the button and the PAD value associated to each word as reported by Mehrabian should correlate. The by-the-users-selected PAD values were logged to a central server as well as the time it took to select the expression. Of these 16 participants, 4 (all male) did not use the wheel even though it was explicitly mentioned. They indicated they did not think of using the wheel in a button. The final set of participants in this first experiment included in the statistical analysis therefore consisted of 12 Dutch persons (5 female, 7 male), aged between 23 and 34, of different occupation (majority not student) and nationality, having normal to good computer skills. The data was statistically analyzed using SPSS 16 and Excel (statistiXL plug-in).

3.4. Experiment 2: Results and Discussion

To evaluate concurrent validity, we correlated the average PAD values (Figure 2) for each emotion as entered by the participants with the PAD values as measured by Mehrabian (1980). This gives a measure of how well the affective feedback correlates with the validated PAD values associated with the word. The average P, A and D values for the different emotions correlated rather well with those found by Mehrabian (1980) (Figure 3; 3D button - Mehrabian correlations). This provides evidence that the button indeed measures the affective quality of the presented emotion word. Interestingly, Pleasure and Dominance values of the original emotion word PAD values correlate with each other (r=0.602, p<0.001). This scale dependency effect is replicated by our study, i.e., our pleasure values correlate with the original dominance values of the emotion words and our dominance values correlate with Mehrabian's pleasure values (Figure 3). This provides even more evidence that the button captured the right construct, i.e., the underlying PAD values of the words as interpreted by the user. In a sense the button has a better validity than the original scoring mechanism used to generate the word's PAD values as our P and D scales are independent from each other (no correlation between 3D AffectButton P and D). This adds to the discriminant validity of the button (the extent to which P, A and D measure different things).

Second, we wanted to know how reliable users score a particular emotion word, i.e., how reliable a measurement tool the AffectButton is with respect to measuring affective attitude towards emotion words. To do so, we exploit the fact that each user is an independent rater. This means we have 12 different P, A and D ratings for each emotion word. The extent to which users agree on the P, A and D values of an emotion word by using the AffectButton can now be expressed by Cronbach's alpha. Cronbach's alpha calculates the average correlation between pairs of ratings, in our case the average correlation of P, A and D (separately) between pairs of users. So, in our case we use Cronbach's alpha as a measure for inter-rater consistency. A high alpha means that users use the button consistently (words are scored the same across users). The alphas for P, A and D are .97, .91 and .95 respectively. As alphas larger than .75 are considered good agreement, this means there is a strong agreement across users when they score the emotion word using the AffectButton. The button has high internal reliability.

Finally, to find out if users learn to use the component, we correlated the time needed to select an expression with the order of presentation of the emotion words. If users learn to use the component, the more words a user scores with the button, the shorter the time needed to do so. If learning occurs, a significant negative correlation should be found. Indeed, users seem to learn to use the AffectButton, as order of presentation and time needed for expression selection correlated with r=-0.256 and significance <0.001. For the first 16 presented words, users needed 12.3 seconds (median) to match a word with an expression, and for the last 16 words users needed 9.8 seconds (median) to select a matching expression (Mann-Whitney, U(192)=22517, p<0.001). We used the median instead of the mean for the following reason. Users were free to do the experiment at home. Although the instructions stated that users should not take breaks, this did happen and therefore we have to account for outliers in the timing data. The median is better suited than the mean in such cases.

A MANOVA with repeated measures was conducted to examine an effect for the measuring tool (independent within-subject variable: AffectButton or Mehrabian) on the standard deviation of emotion words for the three emotion scales P, A and D (dependent variables). The analysis found a significant main effect (F(3, 29)=22.7; p<0.001) for measuring tool. This effect was again found in a univariate analysis on the std of Pleasure (F(1, 31)=13.8, p=0.001) and on the std of Arousal (F(1, 31)=52; p<0.001) with the Arousal scale being by far the strongest contributor to the effect. Further, the AffectButton's Arousal scale had twice as high a standard deviation (0.61) as its Pleasure (0.29) and its Dominance (0.35) scales. These analyses indicate that there is a considerable amount of variation in the resulting Arousal scores. Finally, our Arousal scale correlates with Mehrabian's Pleasure scale. This is not the case for Mehrabian's Arousal scale. These statistical facts together with the fact that 4 out of 16 subjects did not use the wheel at all and that these subjects indicated that they did not think of doing so in a button, lead us to conclude that further testing is needed with regards to the scroll wheel. An additional argument for simplification of the button is that experiment 1 showed that the amount of effort involved in using the AffectButton is perceived as higher than standard rating mechanisms. We address this simplification in the third experiment where we repeat the same experimental setup with a simplified version of the AffectButton.

3.5. Experiment 3: 2D AffectButton without Wheel Control

In this last experiment we mapped the 3D PAD space onto a 2D space, such that the Pleasure, Arousal and Dominance axes can be controlled using just the mouse x and y coordinates in the button. Users do not need to use the mouse wheel anymore. The mapping was done such that at the extremities of the button, the Arousal part was controlled, while at the interior of the button the Pleasure and Dominance where controlled. The details of this mapping are out of scope for this paper, but can be requested from the authors and can be found in the online Java code. The experiment setup was exactly the same as in experiment 2. We contacted 50 subjects, of which 11 participated voluntarily (3 female, 9 male, majority non-student, all Dutch). Also in this task the goal was to find the best matching expression for a word using the AffectButton. The aim in this experiment was again to evaluate the concurrent validity, reliability and usability of the button just as in experiment 2.

3.6. Experiment 3: Results and Discussion

Again, the average Pleasure, Arousal and Dominance values for the different emotions correlated very well with those found by Mehrabian (Figure 3, 2D button – Mehrabian correlations). The size of the correlations were more or less similar to those produced by the 3D AffectButton, indicating there is no performance loss due to flattening the 3D space. Further, Arousal scored using the AffectButton did not correlate anymore with Mehrabian's Pleasure scale, indicating that users feed back Arousal independently from Pleasure, as should be the case. The scale dependency effect between Pleasure and Dominance was again replicated.

We again evaluated the reliability of the data produced by using the button. As in the second experiment, users are assumed to be raters, and the alpha is used as measure of agreement between raters for each emotion word. Alpha was 0.97, 0.94, and 0.96 for Pleasure, Arousal and Dominance respectively. This means that the 2D AffectButton performs comparably well with regards to internal reliability.

Again a MANOVA with repeated measures was conducted to examine an effect for the measuring tool (2D AffectButton and Mehrabian) on the std of three emotion scales. The results indicate a significant main effect (F(3, 29)=3.8; p. = 0.021) for measuring tool. Still, considering the F-ratio the effect seems far less than in the 3D button case. Further univariate analyses showed that the only factor contributing significantly to this effect was the Arousal factor (F(1, 31)=11.7, p<0.002). A t-test confirmed that the standard deviation of the 2D AffectButton's Arousal scale is significantly less than the standard deviation in the Arousal scale of the 3D version (0.42 instead of 0.61, t(31)=4.6, p. < 0.001). This indicates that removing the scroll wheel function was indeed a step into the right direction, as it removed considerable variation from the emotion word scoring making the button more reliable.

Finally, we analyzed the amount of effort involved in scoring emotion words using the button. In the first half, users needed 9.1 seconds to find their best matching expression (median), and in the second half users needed 7.3 seconds (Mann-Whitney, U(176)=18106, p<0.01). This is significantly faster (2.5 seconds) than the first version of the AffectButton (Mann-Whitney, U(176, 192)=20089, p<0.01).

Based on the currently available data, we conclude that the 2D AffectButton is more reliable and easier to use than the 3D version with mouse wheel function.

4. General Discussion

The main contribution of this paper is an important step into the direction of reliable, valid, quick and user-friendly explicit emotion measurement instrument. We have shown that the AffectButton can be used to validly and reliably measure a person's affective attitude towards emotion words. We believe that integrative validation aimed at usability, use in context but also reliability and validity of the data obtained is necessary for broad applicability of measurement tools.

Developing valid digital measurement tools is not a trivial task. The challenge is to develop a tool that has a different form, but aims at getting the same kind of data and level of precision as, questionnaires (e.g., Mehrabian, 1980) or manikins [2, 5]. Properly testing the tool with standard stimulus sets is important before using the measurement tool in practice. This is common in experimental psychology, a discipline embedded in HCI research. We find strong correlations, indicating that the AffectButton is a plausible candidate for affective feedback. Further, our experiment in context (affective holiday rating) shows that users are able to use the button for preference feedback.

Cognitive load is an important issue in affect measurement. Although our results are promising--the time needed to select an expression is not problematic considering first-time use, and users do not perceive the amount of effort as problematic--future work will directly compare different affect measurement methods.

The reason why people currently need some time to select an emotion is that they have to get accustomed to selecting an expression in a button. For all subjects this was the first time ever to be presented with the button as well as the emotion-word matching task. Compare for example the use of emoticons. A first time user has to search for and understand how to read emoticons and still emoticons are by now widely used because they serve an important communication need.

Of course, measuring affective feedback for words and holidays is not the same as measuring mood, or a particular emotional reaction. We are fully aware of the difference between mood, emotion, affect, conscious feeling vs. unconscious emotion, appraisal, attribution, etc. However, for the purpose of this paper, what matters is that we measure the affective quality of a stimulus and that the AffectButton can be used to do so. By doing so, we validate the feedback produced by the button. Further, we have used the button in context and show that users can indeed use the button to generate useful feedback. Future work includes using the button in a more natural setting that elicits actual emotions.

One potential drawback of our way of using a factor-based theory as basis for a measurement tool is that multiple emotions might map to the same PAD triplet. We cannot solve this problem with the AffectButton, as this is a consequence of the chosen underlying psychological model for affect. Still, this is not problematic if one is interested in measuring the affective quality of an object/emotional state/mood/etc, which is what we aim to do. Nonetheless, this 3-dimensional model enables representation and subsequent expression of a large variety of emotions including 5 of the 6 basic ones described by Ekman.

A major benefit of the form of the AffectButton is that, in principle, it can also be used to input mood, as the same three factors underlying the AffectButton can be used to describe mood (Mehrabian, 1980). Future work is needed to investigate this claim. Also, it can be used to record a dynamic change of emotion. For example, a user might indicate an emotion shift from happy to sad during a certain event. This can be captured by the button. This opens up interesting possibilities from a computer human interfacing perspective. For example, when asking a user to rate a content item for recommendation purposes, the user could indicate that in the beginning the item seemed fun while at the end the item was boring. Further, as emotions are often argued to be dynamic instead of static expressions, it is important to see and be able to select the expression change in order to match the user's feeling. As an example of this, we invite the reader to try out the AffectButton and experience the difference between moving slowly along the Pleasure axis (mouse-x) from relaxed (+1P, -1A, +1D; in the button top right with the wheel moved away) to frustrated (+1P, -1A, -1D; in the button top left with the wheel moved away) and vice versa. Somewhere in between relaxed and frustrated an expression is generated that can be interpreted as serious or evil, two completely different emotions, but selectable using the AffectButton's dynamic capacities.

5. Conclusion and Further Work

We conclude that the AffectButton we have proposed enables dynamic affective feedback in a straightforward and easy to use way, without the need for questionnaires, fold-outs of discrete facial expressions or other ways that complicate embedding the device in an interface. Analysis of the data produced in three different experiments showed that subjects can use the AffectButton for affective attitude feedback. Users understand the button and learn to use it better. Further, the button produces valid and reliable data.

Future work includes studies into the generic usability of the button as well as further validity and reliability testing. The small sample size used in experiment 2 and 3 limits generalization of our conclusions to a large population. Therefore, future work should include larger and more diverse user samples (e.g., different cultures). Qualitative evaluation is also planned.

An important next step is to evaluate the button in real-life situations, such as to enter affective feedback to real products and content items (e.g., in the case of an "affective recommendation engine") or in the context of social software. Another important next step in our research is to find out if the button is also suitable for measuring the actual emotional state/mood of a person. Using a device when in a (potentially strong) emotional state is certainly different from using a device to give an affective opinion about something.

6. Acknowledgements

This research is supported by the Dutch Technology Foundation STW, applied science division of NWO and the Technology Program of the Ministry of Economic Affairs. It is part of the Pocket Negotiator project VIVI-project nr 08075. Part of this work has been carried out at the Novay, Enschede. We thank Alina Pommeranz and Gerwin de Haan as well as all participants.

References

[1] Bartneck, C., Reichenbach, J., van Breemen, A. In Your Face, Robot! The Influence of a Character's Embodiment on How Users Perceive Its Emotional Expressions. In: Proc. of Design & Emotion, 2004.

[2] Bradley MM, Lang PJ. Measuring emotion: the Self-Assessment Manikin and the Semantic Differential. Journal of Behav Ther Exp Psychiatry 25: 49-59, 1994.

[3] Breazeal, C. Affective interaction between humans and robots. In: J. Keleman and P. Sosik (eds) Proc. of the ECAL 2001, LNAI 2159: 582-591, 2001

[4] Derks, D., Fischer, A.H., and Bos, A.E.R. The role of emotion in computer-mediated communication: A review. Computers in Human Behavior 24: 766-785, 2008.

[5] Desmet, P. Designing Emotions. Doctoral dissertation, Delft University of Technology, The Netherlands, 2002.

[6] Frijda, N. H., Manstead, A. S. R., & Bem. S. Emotions and Beliefs. Cambridge University Press, 2000.

[7] Graesser, A.C., Chipman, P., Haynes, B.C., and Olney, A., "AutoTutor: an intelligent tutoring system with mixed-initiative dialogue," Education, IEEE Transactions on 48: 612-618, 2005.

[8] Gray, W.D. and Salzman, M.C., Damaged merchandise? A review of experiments that compare usability evaluation methods. Human-computer Interaction, 13: 203–261, 1998.

[9] Höök, K. Affective Loop Experiences – What Are They? In: Proc. of Persuasive 2008: 1-12, 2008.

[10] Isomursu, M., Tahti, M., Vainamo, S., and Kuutti, K. Experimental evaluation of five methods for collecting emotions in field settings with mobile applications. International Journal of Human-Computer Studies 65: 404-418, 2007.

[11] Krijn, M., Emmelkamp, P.M.G., Biemond, R., de Wilde de Ligny, C., Schuemie, and M.J., van der Mast, C.A.P.G.. Treatment of acrophobia in virtual reality: The role of immersion and presence, Behaviour Research and Therapy 42: 229-239, 2004.

[12] Laurans, G., & Desmet, P.M.A.. Using self-confrontation to study user experience: A new approach to the dynamic measurement of emotions while interacting with products. In: P.M.A. Desmet, M.A. Karlsson, and J. van Erp (Eds.), Proc. of Design & Emotion 2006, 2006.

[13] de Lera, E., & Garreta-Domingo, M. Ten Emotion Heuristics: Guidelines for Assessing the User's Affective Dimension Easily and Cost-Effectively. Emotions in HCI workshop, British HCI 2007, 2007.

[14] Lang, P.J., Bradley, M.M., and Cuthbert, B.N. International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-8, University of Florida, Gainesville, FL, 2008.

[15] Lewis, M., Haviland-Jones, J. M., & Barrett, L. F. The handbook of emotion, 3rd Edition. New York: Guilford, 2008.

[16] Mehrabian, A. Basic Dimensions for a General Psychological Theory. OG&H Publisher, 1980.

[17] Mulder, I., and van Vliet, H. In Search of the X-Factor to Develop Experience Measurement Tools. In: Probing Experience: 43–56, 2008.

[18] Ortony, A., Clore G. L., & Collins A. The Cognitive Structure of Emotions. Cambridge University Press, 1988.

[19] Panksepp, J. Affective neuroscience: The foundations of human and animal emotions. Oxford University Press, 1998.

[20] Pantic, M., and Rothkrantz, L.J.M. Automatic Analysis of Facial Expressions: The State of the Art. IEEE Trans. on Pattern Analysis and Machine Intelligence 22: 1424-1445, 2000.

[21] Picard, R.W. Affective Computing. MIT Press, 1997.

[22] Picard, R,W, Klein, J. Computers that recognise and respond to user emotion: theoretical and practical implications. Interacting with Computers 14: 141-169, 2002.

[23] Pouwelse, J.A, Garbacki, P., Wang, J., Bakker, A., Yang, J., Iosup, A., Epema, D.H.J, Reinders, M., van Steen, M.R., and Sips, H.J. TRIBLER: a social-based peer-to-peer system. Concurrency and Computation: Practice and Experience 20: 127—138, 2008.

[24] Rolls, E. T. Précis of The brain and emotion. Behavioral and Brain Sciences, 23: 177-191, 2000.

[25] Russell, J.A. A circumplex model of affect. Journal of Personality and Social Psychology 39: 1161-1178, 1980.

[26] Russell, J.A. Core affect and the psychological construction of emotion. Psychological Review 110: 145-172, 2003.

[27] Sánchez, J.A., Hernández, N.P, Penagos, J.C., and Ostróvskaya, Y. Conveying Mood and Emotion in Instant Messaging by Using a Two-Dimensional Model for Affective States. In: Anais do IHC 2006: 66-72, 2008.

[28] Scherer, K.R., Schorr, A. and Johnstone, T. Appraisal Processes in Emotion: Theory, Methods, Research. Oxford University Press, 2001.