AffectButton: a method for reliable and valid affective self-report

Joost Broekens[1], Willem-Paul Brinkman
Intelligent Interaction Department, TU Delft, Delft, The Netherlands

[1]Corresponding author: joost.broekens@gmail.com

Abstract
In this article we report on a new digital interactive self-report method for the measurement of human affect. The AffectButton (Broekens & Brinkman, 2009) is a button that enables users to provide affective feedback in terms of values on the well-known three affective dimensions of Pleasure (Valence), Arousal and Dominance. The AffectButton is an interface component that functions and looks like a medium-sized button. The button presents one dynamically changing iconic facial expression that changes based on the coordinates of the user's pointer in the button. To give affective feedback the user selects the most appropriate expression by clicking the button, effectively enabling 1-click affective self-report on 3 affective dimensions. Here we analyze 5 previously published studies, and 3 novel large-scale studies (*n*=325, *n*=202, *n*=128). Our results show the reliability, validity, and usability of the button for acquiring three types of affective feedback in various domains. The tested domains are holiday preferences, real-time music annotation, emotion words, and textual situation descriptions (ANET). The types of affective feedback tested are preferences, affect attribution to the previously mentioned stimuli, and self-reported mood. All of the subjects tested were Dutch and aged between 15 and 56 years. We end this article with a discussion of the limitations of the AffectButton and of its relevance to areas including recommender systems, preference elicitation, social computing, online surveys, coaching and tutoring, experimental psychology and psychometrics, content annotation, and game consoles.

1 Introduction

The measurement of human affect can be approached from two fundamentally different angles. First, one can try to automatically infer affective user information by measuring the behavior and physiological signals of that user. This can be called implicit affect measurement, or automatic affect detection, as it tries to derive affective information from the above mentioned data and as such is always indirect; it is an affective interpretation of the (possibly multimodal) signals coming from the user (Zeng et al., 2009). Second, the user can be asked to provide affective information. This can be called explicit affective feedback, or affective self-report, as one asks the user directly to judge his or her emotion, mood or preferences/attitudes and enter the resulting affective value in the system. This can be done in many different ways, see for example (Desmet, 2005; Isomursu et al., 2007), including verbal, pictorial, animation-based, and questionnaire based methods.

There is a long history of the measurement of human affect in the fields of psychology and affective computing. Typical explicit instruments used are self-report questionnaires (Watson et al., 1988), adjectives (Russell, 1980), and the well-known Self-Assessment-Manikin (Bradley & Lang, 1994; Lang, 2008). In addition, a set of implicit mechanisms exist that deduce affect and emotion from different physiological modalities such as heart rate, skin conductance, facial expression and voice (for overviews see (Calvo & D'Mello, 2010; Picard, 1997; Picard & Klein, 2002; Zeng et al., 2009)). Classical affective self-report methods

usually consist of paper or digitized questions that have to be answered on a 5, 7 or 9 point scale. Therefore, these are difficult to embed as part of an interface because they take up a considerable amount of screen space. Further, filling in an affect questionnaire, or filling in picture-based tools such as SAM requires specific instruction about how to use the instrument as well as time to click through the rating scales. On the other hand such mechanisms are often well-validated, and therefore offer meaningful feedback. Automatic affect recognition typically works well in laboratory environments, but currently fails when used as a generic mechanism to capture spontaneously expressed affect. This failure is due to the large number of confounding factors in real world usage (Zeng et al., 2009), such as emotion expression being context and task dependent, lighting and background sounds introducing noise, positioning of the user influencing recognition, and emotion display rules restricting expression (Calvo & D'Mello, 2010). Also the field had a historical focus on the recognition of *acted prototypical affective states* (e.g., extreme anger or happiness expressed on purpose) instead of spontaneously expressed affect. This is a problem, because people typically express a much wider variety of affective states and in more subtle ways than prototypical anger, sadness, etc. As a consequence, many of the methods developed in the past worked well for classifying very clearly expressed emotion, but not for spontaneous, more subtle and varied expressions. Further, the sensors needed for multimodal affect recognition are often not available in real-world settings. This is not to say that automatic affect recognition is not possible or fruitful, but more so to stress that there currently is a need for a digital method for affective self-report that is able to capture human affect in a reliable, valid and user-friendly way, as automatic recognition cannot yet fulfill this need. It is with this in mind that we propose and evaluate the AffectButton.

In this article we focus on explicit affective feedback also known as affective self-report. In Section 2 we argue that measurement of affect is important for human technology interaction (HTI) in general and for affective HTI in particular. Further, we show that there are limitations to the currently available methods for automatic affect detection and affective self-report that limit the gathering of affective feedback. In Section 3 we propose a novel self-report method called the AffectButton. The AffectButton enables users to express affect, in terms of Pleasure, Arousal and Dominance values ranging between -1 and 1, by a single click on a dynamically changing iconic facial expression. *Pleasure* relates to the positiveness versus negativeness of affect, *Arousal* to the level of activation, and *Dominance* to whether the environment is imposing an influence over us or the inverse (Bradley & Lang, 1994; Mehrabian, 1980; Osgood, 1966; Russell & Mehrabian, 1977). Due to its small size the AffectButton is easy to embed in an interface, and, because it presents the user with a facial expression, users do not need an explanation of affective dimensions before using it. The AffectButton can for example be used as an alternative to paper or digital forms of the Self-Assessment-Manikin (Bradley & Lang, 1994), an icon based rating scale with three affective scales (see Fig 1). We claim that the AffectButton can be used as a standard interface component for affective self-report by showing that it is reliable, valid, and usable. In Section 4 we present supportive evidence for this claim in the form of an analysis of 8 experiments with the AffectButton. In Section 5 we discuss the results of the AffectButton, and put these in a theoretical and applied context.

2 Affective HCI, affect and self-report, and challenges

To position our research in the field of affective computing and human-computer interaction and argue for the relevance of the measurement of human affect in general, we first provide a short overview of the use of affect in human-computer interaction. Then we discuss how theoretical views on emotion and affect can influence the design and workings of an affective self-report method. After that, we focus on our main motivation for the development of the AffectButton by reviewing existing affective self-report methods. The aim of this review is to extract a set of challenges that need to be addressed by an affect self-report method.

2.1 The measurement of human affect is crucial in affective human technology interaction

Emotion plays a crucial role in humans and other mammals. The capacity to have emotion is shared among mammals, each species having their own specifics, but all sharing at least emotional states that relate to fear, expectancy, rage, and panic (Panksepp, 1982). Emotions help humans to prepare for action (Frijda, 1988, 2004), to evaluate the situation at hand in a heuristic and efficient way and select appropriate behaviors in a timely manner (Damasio, 1996), to fixate, provide meaning and priority to believes and ideas (Frijda et al., 2000), to reflect upon and learn from past behavior (Baumeister et al., 2007), to communicate intentions to other individuals (Fischer & Manstead, 2008) including individuals from different species such as human-dog communication (Pongrácz et al., 2005), to feel and show empathy towards others (Fischer & Manstead, 2008), and not in the least to give feedback to others (Tunstall & Gipps, 1996) allowing social species to raise their offspring in an effective manner. As such, emotion plays a key role in behavior, cognition and communication.

When humans interact with inanimate technology, we have the tendency to attribute to this technology agency, intentions and an internal life, a well-known and by now classical finding (Reeves & Nass, 1996). It is therefore plausible to assume that emotion also plays a key role in how we, as humans, interact with our technology as well as interact with each other through technology. This view has been advocated by many in the recent past (Bickmore & Picard, 2005; Derks et al., 2008; Höök, 2008; Hudlicka, 2003; Paiva, 2000; Pantic et al., 2005; Picard, 1997, 2003; Picard & Klein, 2002).

Abundant examples are available showing how humans can benefit from using emotion in the interaction with technology including affective pedagogical agents that help students learn skills by giving motivational feedback (D'Mello et al., 2007; Graesser et al., 2005; Woolf et al., 2010), affective agents that play believable opponents in virtual reality negotiation training (Core et al., 2006), empathic agents (Brave et al., 2005; Hone, 2006; McQuiggan & Lester, 2007; Paiva et al., 2004), robots that learn from human emotions (Broekens, 2007), the modeling and measurement of affective user preferences (Broekens, Pommeranz, et al., 2010), affect-adaptive games that adapt to the player in order to enhance gameplay (Gilleade et al., 2005; Hudlicka, 2008; Sykes & Brown, 2003; Yannakakis & Hallam, 2009), and enhanced user acceptance of robots that show social cues (Heerink et al., 2006).

In addition to approaches aimed at using emotion to enhance HCI, there are systems that try to capture human affect to enhance human-to-human communication (Derks et al., 2008), such as measuring arousal (heartbeat) to increase feelings of intimacy over distance (Janssen

et al., 2010), and using emoticons to enhance instant messaging experience (Sánchez et al., 2008).

In most – and possibly all – affective HCI approaches, the final goal is to either measure user affect to do something useful with it (e.g., make a robot learn better) or influence the user's emotion (e.g., create a more exciting gaming experience, or a more intimate chat session). This idea of having this closed loop in which human affect plays a key role in the interaction between the human and its technology has been called the *affective loop*[1] (Conati et al., 2005; D'Mello et al., 2007; Paiva, 2000). Here, affect can refer to feeling, emotion, mood, attitude/preferences, or personality traits (Broekens, 2010). The affective loop usually consists out of a five step process involving different affective abilities (Picard, 1997) of the technology involved, although these steps need not all be instrumented in a particular system to the same level of detail and sophistication:

- user affect detection; the system detects the user's emotion, mood, preference, etc. either using automatic affect recognition or a method for affective self-report. This is then used as data by the system to be interpreted in the usage context. For example, user affect is detected from static faces and classified as one of the six basic emotions proposed by Ekman (Pantic & Rothkrantz, 2004). An example of explicit affective feedback is a user indicating his emotional response to alternative designs of a new car (Desmet, 2005).
- user affect interpretation; the affective data from the user is interpreted in the current interaction context. For example, smiles are interpreted as positive learning feedback while anger expressions are interpreted as negative feedback for an adaptive robot (Broekens, 2007).
- system affective state synthesis; the system can have a representation of its own "synthetic" affective state that can be influenced by the user's affective feedback but also other events detected by the agent. For example, an agent in a virtual reality training changes its emotional state according to recent events in the training caused by the user (Core et al., 2006; Gratch & Marsella, 2001).
- synthetic affective state expression; a system with an affective state typically needs to communicate this to the user using one or more affective modalities (facial expressions, sound, bodily movement, etc.). For example, a tutor agent detects that the user is not making progress and is frustrated, making the agent choose a particular motivational expression to encourage the user (D'Mello et al., 2007).
- influence user affect; the user attributes particular affective meaning to the expressions of the system, thereby influencing the user's affective state, or, the behavior of the system is such that it implicitly influences the user's affective state. An example of the second is a system that communicates heartbeats from one user to another in order to increase feelings of intimacy (Janssen et al., 2010). An example of the first is a system that encourages a user by showing empathic expressions in appropriate situations (McQuiggan & Lester, 2007).

The measurement of human affect plays a crucial role in the first and last of these affective abilities. First and foremost, when detecting affect one needs to measure affect. Second,

[1] Note that the group of Höök (Fagerberg et al., 2004; Höök, 2008; Sundström, 2005) gives the affective loop a more specific meaning, i.e., enhancing *a particular* affective experience using technology to provide a positive feedback loop.

when influencing user affect, the intended effect needs to be evaluated (e.g., does a motivating expression of a virtual agent reduce user frustration). Sometimes this is done at research time, sometimes at run time, but in any case valid and reliable affect measurements are needed. Furthermore, in any setting where one is interested in user affect the measurement of affect is key, even in settings that do not relate to affective loops per se, such as public opinion research, psychological experiments, feedback to consumed media such as through the return channel of interactive television (Lee et al., 2007), recommender systems (Hsu et al., 2007), and product design (Desmet, 2005).

2.2 Affect and how it influences detection and self-report methods

The measurement of human affect is crucial for affective HCI. However, what is affect? In this section, we introduce relevant terminology and three main theoretical perspectives on affect and emotion. We discuss these perspectives because the choice for one or the other strongly influences the nature of an affect measurement method. Finally, we argue that as affective factor values underlie each mood, emotion and attitude, a factor-based approach towards affective self-report is a generically applicable one, though not one that is able to extract detailed emotional differences. It is with this aim in mind that our contribution should be seen, equivalent to the approach underlying the Self-Assessment-Manikin (Bradley & Lang, 1994).

In general, an *emotion* is an adaptive response of an organism in order to cope with a change in the environment and has associated with it a particular facial expression, feeling, cognitive process, 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 and powerful experience. For example, I feel (and notice I am) happy about seeing an old friend. Emotions in psychology are typically labeled, such as anger, fear, guilt and pride.

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. Because moods are typically undifferentiated, these are described in terms of affective factors, such as Valence (positive versus negative) and Arousal (active versus passive), and usually (but not always) lack clear labels in psychology.

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, and I feel *enthusiastic* about theme parks. Because an attitude refers to the associated affect one has with someone or something, attitudes come in a large variety of descriptions including labels, such as exciting, worrisome, and cool, and factors, such as "I like this a bit" referring to how much one values (valence) a particular thing.

To complicate matters a little, *affect* has two different meanings. First, it is used as commonplace term for everything that has to do with the above. We do not refer to this "umbrella" meaning when we talk about measuring affect. Second, affect refers to an underlying core present in emotion, mood and affective attitude towards persons and

things. We use affect to refer to this underlying core, also known as *core affect* (Russell, 2003; Russell & Mehrabian, 1977).

Affect and emotion are complex topics, and agreement on solid definitions does not really exist. The above should be interpreted as descriptions of what the different emotion-related terms usually refer to. We have not tried to formally define emotion or affect , as many excellent works have been published from different perspectives that together do much more credit to the diverse and multimodal nature of emotion (Frijda et al., 2000; LeDoux, 1996; Lewis et al., 2008; Ortony et al., 1988; Panksepp, 1998; Picard, 1997; Rolls, 2000; Scherer et al., 2001). For a quick and broad introduction to the different emotion theories and the history of these see chapter 5 in (Eysenck, 2004) or chapter 3 in (LeDoux, 1996).

To appreciate the theoretical basis of the AffectButton, we do need to explain the factor-based perspective in more depth. To do this, we need to first introduce the three main theoretical perspectives on emotion in psychology. First, there is the categorical perspective, such as the well-known six basic emotions as proposed by (Ekman, 1993). Emotions are split up in categories, e.g., fear, anger, happiness, surprise, disgust, sadness, and each category has a particular expression, feeling, and action tendencies associated with it making it as it is. For example, fear-related emotions prepare for flight and avoidance, feel negative and have a characteristic facial expression. Second, there is the componential perspective focused at explaining emotion elicitation. Cognitive appraisal theories (Ellsworth & Scherer, 2003; Ortony et al., 1988; Scherer et al., 2001) are componential, and describe emotion as a combination of the activation of different evaluative processes addressing personal relevance of an event. For example, anger is the result of an evaluation of an event that was bad for my goals (low goal conduciveness), caused by someone else (agency) who had possibilities to do otherwise (responsibility). For a recent overview and reflection upon cognitive emotion elicitation, see (Gratch et al., 2009). Finally there is the factor-based perspective (Bradley & Lang, 1994; Mehrabian, 1980; Osgood, 1966; Russell, 2003; Russell, 1980) describing emotion in terms of two (Pleasure, Arousal) or three (Dominance) continuous factors. For example, happy is a pleasurable, slightly arousing and dominant experience, while fear is not pleasurable, highly arousing and submissive. In fact, a factor-based perspective can be used to describe the affective component in moods, emotions, and attitudes about things and persons.

These perspectives strongly influence the nature of affect measurement. Choosing a categorical perspective means trying to measure prototypal emotions or attitudes, possibly including the intensity of the prototype. This would result in a self-report method that allows the user to pick one or more emotions and indicate the felt intensity. An example is PrEmo, the animation based self-report method by Desmet (Desmet, 2005). Choosing a componential perspective means trying to elicit the underlying processing components responsible for the emotion. This would result in a self-report method that questions the user about the causal factors of the felt emotion, for example by asking the user to explain (or think aloud while experiencing) the situation. An example that would accommodate for this kind of feedback is the experience sampling method (Isomursu et al., 2007). Choosing a factor-base perspective means trying to extract core affect, i.e., the values on two or three underlying affective dimensions for a mood, emotion or attitude. This would result in a self-report device that either directly or indirectly tries to extract Pleasure, Arousal (and possibly

Dominance) of an emotion, mood or attitude. An example of a direct approach is the Self-Assessment-Manikin (Bradley & Lang, 1994). An example of an indirect approach is the work by Sanchez (Sánchez et al., 2008), and the AffectButton presented and analyzed in this article.

These perspectives also strongly influence the capabilities and limitations of an affect measurement method. A categorical approach would by definition have trouble extracting moods, as moods are not categorical. It would also have to address (by ignoring or making explicit use of) the fact that users can feel a mixture of different emotions. For example, one can feel both happy and sad when a great party has ended. Is this feeling a special category? Is it composed of two active categories at the same time? Is the intensity of both lower because they are both active? On the other hand, a categorical approach can measure specific emotional differences, such as guilt versus sadness, simply by asking (verbal tool) or showing the emotion (pictorial-, or animation-based representation). A componential approach will have difficulties automatically integrating measurements across people, as the data gathered is very subject specific. Also, measuring mood in a consistent way will be difficult, as mood is typically not related to a particular event and people are bad at identifying the causal factors of their moods. On the other hand this approach can measure subtle differences such as the difference between shame or guilt (very hard to distinguish based on a picture, and impossible to distinguish based on factor-values). A factor-based approach would be unable to measure differences between specific emotions that share the same factor values, such as guilt versus shame or happy-for versus happy-about. The factor values for guilt and shame are almost identical, just as for happy-for and happy-about, but have a different emotional meaning. On the other hand this approach is applicable to mood, emotion and attitude, as these three affective phenomena share the fact that they are expressible in a factor-based system, as explained now.

Disregarding these different theoretical views, most emotion researchers believe that there are two common affective factors that are useful to describe a mood, emotion or attitude: *Valence* (*Pleasure*) 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. This essentially means that factor values underlie each mood, emotion and attitude, making a factor-based approach a generically applicable one, though not one that is able to extract detailed emotional differences. As we aim for a generic digital affective self-report method, it is with this aim in mind that our contribution must be seen.

2.3 Challenges for affective self-report methods

We have argued that there is the need for a generic digital affective self-report method. Now, we review existing self-report methods and identify five main challenges that such a tool must address: the tool must produce reliable and valid data, and must be usable, generic, and embeddable in interfaces. These challenges have been our motivation for the development and evaluation of the AffectButton.

Several approaches exist towards explicitly gathering affective feedback by means of digital methods (Desmet, 2002; Isomursu et al., 2007; Sánchez et al., 2008). Typically, these

approaches have a fundamental tradeoff between precision and measurement speed/ease of use (Isomursu et al., 2007). 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 providing affective feedback. In our approach we specifically aim for both precision and speed/ease of use (but not emotional specificity, as a factor-based approach is unable to address this criterion as discussed above).

Approaches that involve high detail and effort include those that focus on extracting detailed affective information by enabling people to elaborate on their appraisal of an object. This typically involves the use of human observers to evaluate the feedback and is focused on measuring human emotion and product attitude during the process of product development (de Lera & Garreta-Domingo, 2007; Isomursu et al., 2004; Laurans & Desmet, 2006). Further, there are approaches that use physical objects as feedback mechanism, where the user has to select an object that best matches his or her emotion or attitude (Isbister et al., 2007; Isbister et al., 2006; Laaksolahti et al., 2009). Such objects make use of abstract affective principles, such as round versus spiky forms, and small versus large forms, to manipulate the affective association people have with the object. As a result, people can consistently match objects with their felt affective experience. Usually the object picked will be the start for a discussion between the researcher and subject to further elicit the experience. Both types of approaches extract detailed affective feedback, but effort is involved in both rating as well as interpretation of the data. Further, the data is not immediately suitable for computation. Our goal is the latter. We aim for a simple, easy to use method to gather affective feedback without the need for human observer intervention.

Further, many methods are based on categorical feedback. The user is asked to provide feedback in terms of categorical (discrete) emotions, such as happy, sad, angry, jealous, with or without intensity, using pictures or animations. The methods focus on extracting user preferences or affect attributed to products or settings. For an overview see, e.g., (Desmet, 2005). The benefit is that discrete emotions are easy to interpret by the user giving the feedback and the person or system interpreting the feedback. This can be particularly important for young children, such as in the development of Sorémo, a tool to measure young children's emotions during the learning process (Girard & Johnson, 2009). Another benefit is that specific emotional differences can be measured better as discussed earlier. A drawback is that moods and in fact anything containing core affect but not a *specific* emotion, including situational descriptions, pictures, and sounds, are difficult to express as these do not come in clear categories. Further, mixed emotions are difficult to interpret as there is no logical "emotional continuum" between categories, and, inputting mixed emotions costs more effort from the user because picking two emotion categories requires more interaction than picking one. Our approach focuses on gathering feedback on affective dimensions. Affective dimensions define an affective space. The affective feedback consists of a point in this space. The benefit of a dimensional approach is that it enables graded, by which we mean of different intensities, and mixed affective feedback that is easy to interpret numerically by a computing device[2]. Factor-based feedback enables numerical operations such as averaging and clustering of affective feedback. This is important for computing systems, for example in recommender systems because this allows the computation of average affective profiles for pieces of content or affective distance between user profiles.

[2] See the discussion for a more detailed statement on mixed emotions.

Finally, several methods exist that are based on a factor-based, i.e., dimensional, perspective. Key in these methods is that the goal is to extract affective feedback in terms of values on affective dimensions, such as Pleasure, Arousal and Dominance. For example the Self-Assessment Manikin (SAM) (Bradley & Lang, 1994) allows the user to rate affect for Pleasure, Arousal and Dominance. The user does so by selecting for each factor a picture from a set of pictures showing emotional faces that express different intensities for that factor (Fig. 1). An example of a target-group specific instrument measuring only the valence dimension is the Smileyometer (Read, 2008), an instrument for measuring children's opinions by using a 5-point Likert scale consisting of smileys ranging from bad to good. Although the SAM method is by now well-validated, a potentially irresolvable limitation is that users must be explained the three affective dimensions before they can use the method. Instruction is even required as defined in the SAM usage protocol. Also, the dominance dimension is notoriously difficult to explain and use because it is defined by a relation between the rater and the environment (one is dominant *with respect to* the dominance of something else) (Broekens, 2012). Finally, if a user is presented with SAM for the first time, and the only instruction is "use this to rate how you feel", that user will have a hard time figuring out that an explosion in the stomach means "being aroused". The only scale that intuitively makes sense from a conceptual and rating point of view is Valence (positive versus negative facial expression). The other scales, as well as the fact that the scales are to be interpreted as one coupled rating, need considerable explanation. This hinders adoption of precise affect measurement outside the domain of research (and indeed we do not know of any user-oriented service that uses SAM or a derivative), and hinders usability with those groups that do not cognitively understand these dimensions yet, such as children. A second drawback is that the method takes up a considerable amount of screen space, and is thus difficult to embed in an interface. This is particularly relevant for mobile devices, or embedded technology with small screens. A third drawback is that the user needs to click three times in order to give feedback. The AffectButton aims to address all three issues. In essence we developed a digital affective self-report method that requires limited user effort and no user interpretation of the three dimensions. As such our work is related to FEELTRACE (Cowie et al., 2000) and the Affect Grid (Russell et al., 1989), with two key differences. First, the AffectButton enables affective feedback on three dimensions analogous to SAM, while FEELTRACE and Affect Grid use two (Activation/Arousal and Evaluation/Valence). Second, the AffectButton shows the user a state representation that represents the currently selected values on the three dimensions in the form of a facial expression, eliminating the need for users to understand the semantics of, and relation between affective dimensions. Of course the user still needs to understand how to use the complete input range of the AffectButton, and be able to attribute affect to facial expressions. A large part of the validation studies is aimed at understanding usability. In the discussion we elaborate on the theoretical basis for attributing affect to faces, in line with (Russell, 1994).

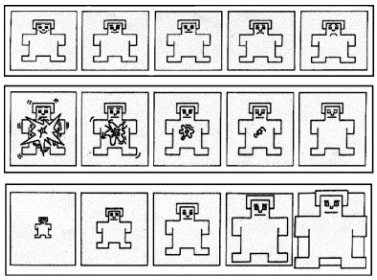

Figure 1. Self-Assessment-Manikin. From top to bottom, the pictorial scale for Valence, Arousal, and Dominance. Normally subjects rate a stimulus (or their feeling) on a 9-point scale with the points being on, or in between two, pictures.

2.4 Summary

In this review we have argued the following. First, the measurement of human affect is important for human technology interaction in general and for affective HTI in particular. Second, there are limitations to current versions of automated and self-report methods of affect measurement. Third, the challenges for an affective self-report method are validity, reliability, and usable, as well as that it should be easy to embed in an interface, that the data is machine interpretable, and that it is generic. In the remainder of this article we introduce the AffectButton and evaluate its usability, reliability and validity.

3 The AffectButton

We now give a short description of the AffectButton (Fig. 2). The AffectButton was developed as a simple button (currently implemented in Java, Python and HTML5 available from http://www.joostbroekens.com). It is a square and is scalable in size, although the evaluation studies presented in this paper all use the standard size (100x100 pixels). It does not unfold or pop-out, and can therefore be considered a static element in an interface. The mouse *x* and *y* coordinates within the button define the values on the affective dimensions Pleasure, Arousal and Dominance. So, a mapping exists from [*x*, *y*] to [*P*, *A*, *D*]. The resulting values are numbers between -1 and 1 on each affective dimension. Affective value triplets are represented by a dynamically changing rendered facial expression. The user can therefore select a large range of affective values from the *PAD* spectrum by moving the mouse within the button, looking at the resulting expression, and clicking. The expression changes while moving over the button. Therefore, the user does not need to interpret affective factors in order to use the button. The user needs to match his/her felt affect with the affect present in the facial expression. For example, a user can move the mouse to the lower-left part. The face will start to express fear (Fig 2, down, third row, right end). When the user clicks the AffectButton it will output a triplet similar to (-1, 1, -1) reflecting negative, aroused and submissive affect. When the user now moves towards the center from the last location the face will gradually change into an expression of sadness. If the user clicks on, let's say the 8th face of the third row (Fig. 2, below), i.e., in between sad and afraid, the AffectButton will output the triplet (-.7, 0, -.7). This could indicate sadness but with a bit of worry. We now discuss the AffectButton's workings and rationale for its design[3].

[3] Please see Section 5 for an elaborate discussion of the AffectButtons workings related to theoretical aspects such as orthogonality of the dimensions, alternative [*x*, *y*] to [*P*, *A*, *D*] mapping, the relation to SAM, and the relation between dominance and core affect

3.1 Pleasure Arousal Dominance as basis

A main motivation for our research has been to develop a generic measurement device for affect, regardless the type (attitude, emotion, and mood). Therefore, we have chosen a dimensional approach that relates to core affect as explained above. As a substantial number of emotions cannot be represented clearly as different points in the Valence-Arousal 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* (i.e. *control*) (Bradley & Lang, 1994; Mehrabian, 1980; Osgood, 1966) (*PAD*), and is therefore more expressive than the 2 dimensional. In this theory *P* relates to the positiveness versus negativeness of affect; *A* to the level of activation, and *D* to whether the environment is imposing influence over us or the inverse. For example the difference between anger and fear can now be represented in a logical way: anger is a negatively arousing dominant emotion, fear is a negatively arousing submissive emotion. The *PAD* factor-based theory states that every object/emotion/mood/etc has a mapping to a point in *PAD* space. The reverse mapping is not the case, i.e., not every point has a unique emotion attached to it. The mapping is many (object/mood/emotion) to one (*PAD* triplet).

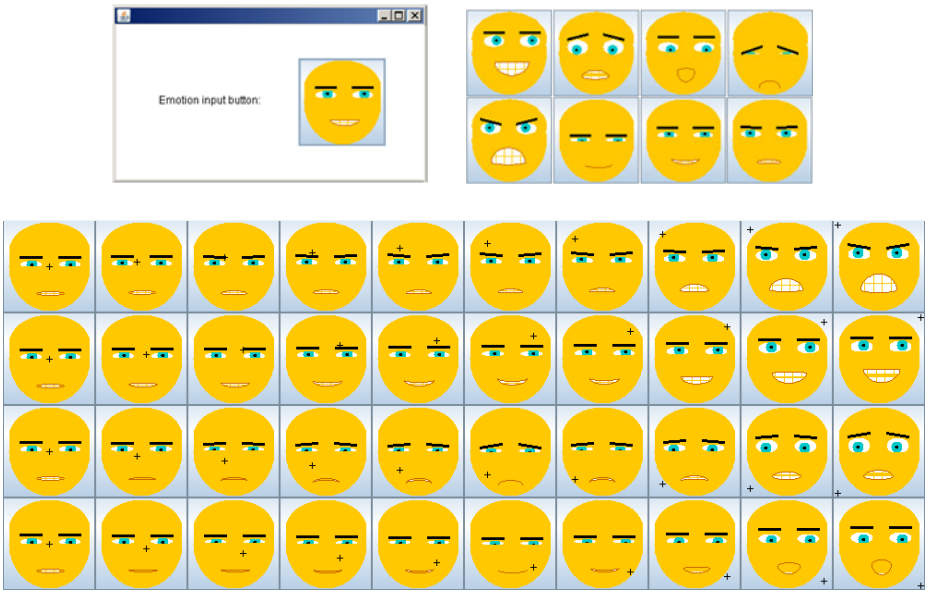

Figure 2. *The AffectButton in a simple window (left), its extreme affective states (right), and four example trajectories from neutral to extreme PAD states with corresponding mouse pointer location (down). Extreme states are elated (PAD=1,1,1), afraid (-1,1,-1), surprised (1,1,-1), sad (-1,-1,-1), angry (-1,1,1), relaxed (1,-1,-1), content (1,-1,1), frustrated (-1,-1,1). Labels are exemplary. Note that the AffectButton allows for continuous input in the PAD space, the extreme prototypes and trajectories are only depicted in this article to give an idea of the complete affective space covered by the AffectButton.*

3.2 Mouse coordinates to *PAD* mapping

Even though we use three dimensions to represent affect, the AffectButton only affords 2-dimensional input by moving the mouse in the button plane. We started with a version of the AffectButton that afforded 3D input (the third dimension was controlled by the mouse wheel), but this proved to be too difficult to use (Broekens & Brinkman, 2009). As such we decided to drop the independence of Arousal (see also Sections 5.2.3 and 4.5). The 2D

mouse coordinates are mapped to *PAD* coordinates in the following way (see also Fig. 3). The button consists of 3 parts, a center part, an inner border and an outer border. In the center part the user only controls *P* and *D*, with the center point being [0, 0]. In the inner border the user also controls *A*, with *A* being interpolated from -1.0 (start of inner border) to 1.0 (start of outer border) based on its distance to the outer border. In the outer border *A* is always 1.0, while *P* and *D* are mapped to their nearest point on the inner border. In essence the outer border does not do anything, apart from providing space for the user to move so that the user will not move outside the AffectButton when trying to express extreme affect. In early studies, including *WordRating1* (see section 4), without this outer border users commented on the fact that they found it hard to select extreme emotions, because they would exit the button altogether while attempting to move the mouse to the edge.

The rationale for this mapping is as follows. High arousal affect is usually associated with what can be described in layman terms as “extreme” emotions, such as rage, enthusiasm, terror, and amazement, while low arousing emotions are usually associated with less ”extreme” emotions such as irritation, happy, sadness, and interest. Our mapping thus matches the intuition of users, and provides an affordance that is easy to understand and is consistent across the button: extreme affect is at the edge and involves more movement of the mouse. Please note that the fact that extreme affects are at the extremities of the button is not the same as claiming that arousal is the intensity of affect, as the quality of affect also changes at these extremities (e.g., sad versus afraid). As such, the design of the button is consistent with (Reisenzein, 1994) claiming that arousal is a factor that has a qualitative influence on affect, and not simply an intensity modifier.

Dominance maps to the Y-axis. *D* relates to body posture; high *D* is associated with open (Cashdan, 1998) and erect (Weisfeld & Beresford, 1982) body postures as well as bigger size (Brown & Maurer, 1986), while low *D* is associated with closed body posture and smaller size. Also, gaze is associated with social control (Kleinke, 1986). This made us decide to map *D* to the Y-axis, as moving up or down with the mouse in this case mimics the behavior of the user’s body to straighten versus slump over. Further, as the AffectButton’s face always looks at the mouse pointer, moving up creates the impression that the face becomes more open and looks the user in the eye, creating a more dominant posture, while moving down makes the face look down and seem to tilt downwards creating a submissive posture. As a result of mapping *D* to the *Y*-axis and *A* control to the extremities of the AffectButton, *P* is mapped to the *X*-axis, i.e., the horizontal axe. This is analogous to Likert-scale ratings or stars, typically used to express preferences on a horizontal scale with positive values at the right side. As such, *P* control on the *X*-axis is an intuitive choice for users.

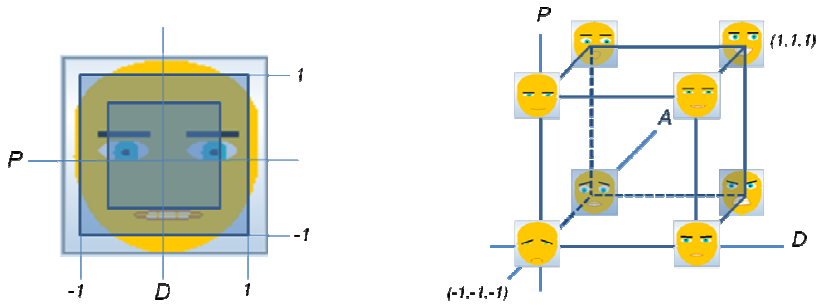

Figure 3. Schematic of the 2D coordinate mapping to PAD space (left), and the location of the extreme prototypes in PAD space that are used for interpolation.

3.3 Rendering the AffectButton's face

Once the *PAD* coordinate is known based on the previous mapping, the face is rendered based on this coordinate using a mechanism that is comparable to the one used in the head of the robot Kismet (Breazeal, 2003). 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 expression). Five of the 8 extremes in the *PAD* space map to Ekman's basic emotions (Ekman, 1993; Ekman & Friesen, 2003) 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 (Mehrabian, 1980; Mehrabian & Russell, 1974). Based on the *PAD* coordinates, the face displayed is interpolated between these nine prototypical expressions.

Each facial expression is defined in terms of eyebrow, eye and mouth configuration of the face, inspired by the Facial Action Coding System (Ekman & Friesen, 2003). These facial features represent the minimal set needed to express all 8 extreme emotions present in the *PAD* affect space. These features are the same as those used for the iCat robot (Bartneck et al., 2004). In addition to that, the face tilts up for high *D* values and down for low *D* values (see rationale in previous section), because it looks at the mouse pointer. As the face looks at the pointer immediately changing facial feedback, it is clear to the user that the button affords moving within its boundaries. The actual expression is based on an interpolation of the features as defined by the 8 prototypical emotions mentioned above (Fig. 2).

To evaluate a bare-minimum button that can still reliably and validly measure affect, the face is rudimentary and gender neutral. 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 style features later. Such an approach gives a clearer view on how such features impact reliability, validity and usability.

4 Assessing the AffectButton's reliability, validity, and usability

In this section we present a comparative analysis of previously published and three novel large-scale *unpublished* studies on the main evaluation criteria reliability, validity, and usability. We discuss and compare the different studies as well as use meta-analysis techniques (DeCoster, 2004) to aggregate statistical findings about reliability, validity and usability across the 8 different studies with the AffectButton totaling 776 subjects. Typically meta-analyses are done with larger number of studies. However, we have used meta-analysis techniques so that we can, in addition to the results of the individual studies, present a coherent overview of the three evaluation criteria. We first give an overview of the different studies that form the basis of this analysis. Each study is described in short so that it is clear what the goal and experimental protocol was, as well as the demographics of the subjects that participated. All experimental results will be discussed in context of the three evaluation criteria presented next in more detail.

Reliability

Reliability is about measurement consistency. If a person indicates he/she is moderately happy, then the next time that user indicates that same emotion it should indeed be comparable, and the same should hold for two different persons. Reliability is important, as

it tells us something about the consistency of the measurement tool (Coolican, 1990). Reliability is addressed in several of the studies presented in this paper, by looking at the level of agreement between users with respect to rating stimuli (*inter-rater reliability*).

Validity

Validity refers to measuring what one wants to measure, i.e., if a user inputs he/she is moderately happy, then this should indeed be the case and not extremely happy or even sad. In our analysis we address several types of validity, and refer to the naming as introduced in (Coolican, 1990). We include *concurrent validity* (relation between measurements with the AffectButton and SAM, both measuring affect at the same time), *predictive validity* (relation between measurements with the AffectButton and existing valid measurements), *convergent construct validity* (relation between measurements with the AffectButton and measurements of other constructs theoretically predicted to be related to affect), and *ecological and population validity* (generalization of AffectButton results with respect to usage settings and users). Convergent construct validity, ecological validity and population validity thus relate to our aim for a generic measurement instrument that is usable for multiple types of affective feedback in various usage settings for a variety of users.

Usability and willingness to use

Another main concern for affective self-report is that it should be useable without (or with very limited) explanation and without supervision, that users should be willing to use the tool, and that users should be able to learn getting better at using the tool. Usability and willingness to use are evaluated in terms of perceived ease of use, perceived liking of the rating instrument, and time needed for rating.

Table 1.
Overview of the different studies.

Study	Stimulus type	Affect type	Sample	n
WordRating1,2	Emotion words	Attribution	Students/Relatives	11, 9
PreferenceRating	Holidays	Preference	Students	32
MusicRating	Film Music	Attribution	High school students	21
AvatarEmotionRating	Written primer	Attribution	Students	48
ANET_SAM1	Affective Texts	Attribution	Dutch women	202
ANET_SAM2	Affective Texts	Attribution	Dutch population young	325
Induction_SAM	Mood Induction/imagination	Feeling	Dutch population average	128
				776

To facilitate reading the analysis, the studies have been labeled, and an overview is given in Table 1. The overall setup of all of the experiments was the same. Subjects were presented with a stimulus or set of stimuli, after which they were asked to rate affect using the AffectButton and in some studies also using SAM. In all studies except the last, subjects were asked to rate affect associated with the stimulus. In this last study, subjects were asked to rate their own mood. In all of the studies, users were explained either orally or in writing how to use the AffectButton (but neither how it works nor the meaning of affective dimensions) and/or SAM (including what the dimensions are, as this is required in the protocol when using SAM). The AffectButton instruction of use consisted (approximately) of the following information: “With this button you can enter what you feel about a [situation|word|etc.]. You move inside the button with the mouse, until you find a face that best matches your feeling. Extreme feelings are on the edge of the button. Please play

around with the button, until you have a good idea about the different possibilities". A critical part of the instruction was the "playing around". Subjects could familiarize themselves with the instrument(s) for several minutes before starting the experiment. No problems were reported with understanding their use, apart from frequent remarks about uncertainty about the meaning of *D* in SAM, and initial uncertainty about the dynamics of the AffectButton. Both problems typically resolved after a couple of trials. All studies used the standard resolution of the AffectButton being 100x100 pixels. Readers primarily interested in the results of the analysis can skip the description of the studies in Section 4.1 and refer to them when needed.

We now give a short overview of why these studies were done. Because we aim for an affect measurement instrument with ecological and population validity (we want to be able to generalize our results), the studies were set up to test affect measurement in a variety of settings. Between these studies the *type of affect* measured, the *subject sample*, and the *type of stimulus given* varies (see Table 1). Type of affect refers to whether we asked subjects to rate their preference, their attribution, or their own mood. It is important to show that the AffectButton can be used to gather various types of affective feedback because affective HCI applications need to know different things from users. For example, a music recommender system needs to know a user's preferences, while an intelligent tutor system that aims to learn math skills to teenagers needs to know the user's mood to adapt the learning session. The subject samples are quite diverse between different studies. This diversity between samples was by design, but the samples themselves where by opportunity. The variation in subject sample enables us to investigate if the AffectButton can be used by a diverse set of people in terms of background, schooling and age. We have a diverse set of subjects, although all Dutch and aged between 15 and 56 years (see discussion). Finally, the variation in stimulus type enables us to investigate if the AffectButton can be used to measure affect associated with different types of stimuli. This is important because affect can be attributed to a wide range of different stimuli such as music, pictures, stories and products. For reliability analysis (Section 4.1) we used the WordRating and and ANET studies. For concurrent and predictive validity (Section 4.2) we used the WordRating and ANET studies. For construct validity (Section 4.2) we used the PreferenceRating, MusicRating, AvatarEmotionRating, and Induction_SAM studies. For usability (Section 4.3) we used all studies except the AvatarEmotionRating study.

4.1 Experiments

1. WordRating1. This study aimed to investigate predictive validity, reliability and usability of the AffectButton (without the outer border that was later added for extra space to facilitate interaction at the button's extremities) by studying how subjects rate affect attributed to emotion words using the AffectButton. It presented subjects with a set of 32 affect words. Each word had valid *PAD* values, and was preselected from a larger list of about 150 words (Mehrabian, 1980) based on two criteria. First the word had low standard deviations on P, A, and D values. Second, the words together spanned the complete *PAD* space as well as possible. As emotion word translations can be problematic, each word was presented in translated Dutch *and* original English (happy/blij) to the subjects in random order. Each subject rated all 32 words. A total of 11 subjects participated (3 female, 8 male, majority non-student and Dutch, University trained, aged 25 - 55 years). Subjects were instructed to read, and rate the word using the AffectButton. We further collected the time needed to

rate the stimulus as objective measure for usability. This study has been published (Broekens & Brinkman, 2009).

2. WordRating2. A replication of WordRating1, but now with the outer border as explained earlier. This study was mainly done to investigate potential effects on reliability and validity due to the addition of the outer border. The setup is the same. The number of subjects equals 9 (3 female, 6 male, majority non-student, Dutch, aged 25 - 36 years). As we found no significant differences in terms of reliability and validity between these two studies (see below), all of the other studies have been done using the AffectButton with the outer border. As the main goal was to replicate findings in *WordRating1*, this study has not been published separately.

3. PreferenceRating. This study investigated whether the AffectButton can be used for rating preferences, instead of simply rating an emotion word having a predefined affective attribution. Main aims were to assess if preferences elicited with the AffectButton are meaningful (convergent construct validity), and to assess the usability of the button in a preference elicitation 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 participants (one had used the AffectButton before, 10 female, 22 male, mainly students and researchers within the field of computer science, Dutch, aged between 21 and 31 years). Each participant had to do both rating tasks. The order of the 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. This study has been published previously (Broekens & Brinkman, 2009; Pommeranz et al., 2008).

4. MusicRating. This study investigates whether high school students can use the AffectButton for the affective annotation of music in real time. The main aim was threefold: investigate if the button is usable by a younger age group, whether it is usable in a different realistic time-critical affect attribution domain like music annotation, and whether the data are meaningful (construct validity). The setup was as follows. The music consisted of three songs from the movie *Pirates of the Caribbean*. These songs were selected for having strong emotional content. Users were told that they were about to listen to 3 songs, that they should use the AffectButton whenever they wanted, but in any case whenever they thought the affective content of the music changed significantly. Then subjects listened to each of the songs, and rated the music in real time. The songs were only listened to on YouTube, no visuals were present. We do not know about preexisting preferences for these songs, but the songs were rated for affective content throughout the song, and participants were asked to rate continuously while listening but in any case when the emotion *in* the music changed. So, preferences probably played only a minor role. However personal interpretation of affect in these songs might play a role as is always the case with any stimulus set. In total 21 subjects participated (18 high school students, 3 parents, 13 female, average student age = 16.7 with *SD* = 1.5, average parents age = 54.6 with *SD* = 3.8). The feedback was time stamped. This resulted in an affect trace for each participant and each song for the duration of a song for

the three affective dimensions of *P*, *A* and *D*. This study has been published previously and details can be found there (Broekens, Pronker, et al., 2010).

5. AvatarEmotionRating. In this study we investigated whether the AffectButton can be used to affectively rate how a Virtual Agent would feel in a situation that was either dominant or submissive. The main aims were to investigate if the AffectButton could be used in another affect attribution domain (a domain that is relevant to the development and evaluation of social agents and robots in particular), and to find out if measured *D* reacts as predicted (construct validity). The setup was as follows. Subjects (*n*=48, 13 female, 35 male, average age = 27, *SD* = 10) had to play a job contract negotiation with an agent. Subjects were first presented with a written primer that either explained that the agent was in a submissive position because it needed to find work and you as a subject were the future boss having a lot of options for filling up the current position (the agent = employee condition), or vice-versa (the agent = boss condition). After this manipulation (between subject), subjects were asked to rate using the AffectButton how they thought the agent would feel. As such, the experimental manipulation was a written primer, not a virtual agent. The virtual agent was presented as context. Parts of this study have been published (Ham & Broekens, 2011).

6. ANET_SAM1. Up until now, our studies involved relatively small sample sizes and limited sample diversity (mostly students, colleagues, friends and family). Also, we have not addressed concurrent validity. In order to address this, we performed three major studies with the AffectButton. This study and the next address in a structured manner the concurrent validity of the AffectButton by comparing AffectButton ratings to SAM ratings and predictive validity by comparing to valid SAM rated situation descriptions (ANET) (Bradley & Lang, 2007). In addition to that, we asked subjects to rate usability factors and timed the average time needed to use the AffectButton and SAM. The setup was as follows. Subjects were recruited and tested at a national women's event (*n*=202, 97.5% female, average age = 43 with SD = 13, 22.3% working full time, 43.0% working part time, 13.4% students and 21.3% unemployed or other occupation, Dutch origin). Subjects were selected on a voluntary basis and were not aware of the actual goal of the study. Subjects were randomly presented with either the AffectButton or SAM as rating instrument (between subject). Of the 202 subjects 102 used SAM while 100 used the AffectButton. Each subject was presented 10 ANET situations (set A) selected at random from the complete set of 60 texts. The random selection was balanced, so that on average each text was selected the same number of times. This resulted in about 17 ratings per text per instrument. We further collected time needed to rate the stimulus, and answers on four 7-point likert scale questions about perceived ease of use, pleasantness, ease of understanding the interface, and effort. This study has not been published yet.

7. ANET_SAM2. Because the ANET_SAM1 study was done with primarily women, we repeated the study with a completely different and more representative sample. The study setup is exactly the same as the previous one, but now subjects were recruited and tested at a culture festival. Subjects were selected on a voluntary basis, and were not aware of the actual nature of the study. Subjects appeared to be a representative sample of young Dutch population (*n*=325, age mean = 27 and SD = 8, 57% male, 37% living in an old urban area, 25% countryside, 38% new urban area, 57% progressive voters, Dutch origin). A total of 164 participants used the AffectButton, while 161 used SAM (again a between subject setup).

This resulted in about 27 ratings per ANET text. This study is also about a comparison between SAM and the AffectButton, as is the previous study. However, due to space/practical reasons we could not use a mouse at the festival where the data was collected, so we had to use the laptop touchpad. This is unfortunate, as this means that differences *between* the two ANET_SAM studies cannot be attributed to input device or sample group. This study has not been published yet.

8. Induction_SAM. This study investigated if the AffectButton is also suitable for measuring actual user mood (not only affect attribution to external items such as words, textual situation descriptors and music). This is needed for ecological validity of the AffectButton because we aim for a generic method for the measurement of human affect. Main aim for this study was to test if a positive versus a stress mood induction (between subject) would influence subject mood ratings as measured with the AffectButton, and to collect qualitative data for the reasons why users preferred either SAM or the AffectButton. The setup was as follows. We performed a large scale online mood induction study. The induction (a 3-minute video) was developed by a professional trainer/coach experienced in treating light forms of stress. The subject was instructed and guided to imagine a recent situation in which they felt stressed (or pleasurable in the positive induction). The only difference between the videos was the instruction to focus on a recent pleasurable versus stressing situation, the technique used for the induction was the same in both. We measured the mood induction using SAM and AffectButton (within subject, order of instrument was random per subject, separate screens). Induction was successful as evidenced by SAM measures (not reported here). A total of 128 subjects participated, recruited from the database of contacts of the coach who developed the induction (71% female, average age = 41, *SD* = 12, education 24% university degree, 41% professional bachelor, 35% lower professional or high school, Dutch origin). Of these subjects 61 received a positive mood induction and 67 a negative induction. First, subjects received a mood induction. Then, subjects rated their current feeling (SAM and AffectButton). Finally, subjects were asked to give feedback on their preferred instrument (forced choice preference, a 5-point scale preference, and a reason for the forced choice). This study has not been published yet.

4.2 Reliability

We now proceed with the analysis of the results of these studies in light of reliability, validity and usability. For each of these criteria, we first explain our method of analysis, after which the results are presented. Reliability was assessed by looking at average rater-total correlations extracted from the different studies[4]. An individual rater-total correlation measures the extent to which that particular rater (subject in our case) agrees with how all other raters in that study rated the stimuli. So, we first calculated for each rater the correlation between his/her item ratings and the average item ratings for all raters. This was done for *P*, *A* and *D* separately resulting in three correlations per subject. Then we calculated the mean correlation over all *n* raters. This average rater-total correlation for *P*, *A* and *D* separately therefore expresses the overall agreement in the subject sample for the different affective dimensions *P*, *A* and *D*. This average is in essence the same as Cronbach's alfa is for

[4] Rater-total correlations were used instead of inter-rater correlations. Both are measures of agreement, but inter-rater assumes considerable overlap in rated items to be able to calculate rater-rate correlations. As raters rated different items in the ANET studies this introduces many missing values for inter-rater correlations. Hence we used the rater-total measure of agreement.

item reliability, but instead of calculating the average item-total correlation over all items, we calculate the average rater-total correlation over all raters. This means we can take Cronbach's alfa criterion for acceptability of rater-total correlations. This criterion sets 0.7 as a target for acceptable reliability, with 0.6 as a lower bound for acceptance (Loewenthal, 2001). As we have 4 different studies, 2 for emotion word stimuli, and 2 for ANET textual situation stimuli, we aggregate first per stimulus type and then for the total. For comparison we also report the reliability for ratings collected with SAM in the two ANET text studies. Within one type of stimulus, the reliabilities are weighted according to the number of subjects (because that is the basis for the average rater-total correlation). Between the two types of stimuli, a simple average is taken. We did not weight between types of affective stimuli because we did not want to give an arbitrary bias towards a particular type of stimulus when assessing reliability simply because the sample of that study is larger. Normally, one would weight according to number of subjects, but in our case we compare different effects (i.e., two different types of stimuli: textual descriptions of situations versus emotion words). Weighting based on sample size would implicitly assume sample is more important than stimulus type, which seems false. Hence we opted for no weighting.

As can be observed (Table 2) the overall reliability for *A* as measured with the AffectButton is 0.66, indicating acceptable reliability, and for *P* and *D* it is 0.81 and 0.77 respectively, indicating above target reliability. When we compare the AffectButton reliability with the reliability of SAM (based on the ANET studies weighted means), we see that the reliability for the AffectButton *D* scale is higher ($t(499)=2.04$, $p=0.042$), while SAM *P* and *A* are higher ($t(519)=-8.34$, $t(499)=-6.24$, $p<0.001$)[5]. Further, the AffectButton's reliability is lower for the ANET studies than for the word rating studies, indicating that type of stimulus and/or sample might influence the reliability of the AffectButton.

Table 2.
Reliability as mean rater-total correlations.

		Self Assessment Manikin			AffectButton		
Study	n	*P*	*A*	*D*	*P*	*A*	*D*
Words							
WordRating1	11				0.89	0.78	0.83
WordRating2	9				0.91	0.73	0.85
Weighted mean	20				*0.90*	*0.76*	*0.84*
ANET texts							
ANET_SAM1	202	0.85	0.67	0.68	0.68	0.56	0.70
ANET_SAM2	325	0.85	0.67	0.65	0.74	0.55	0.71
Weighted mean	*527*	*0.85***	*0.67***	*0.66**	*0.72***	*0.55***	*0.71**
Mean					***0.81***	***0.66***	***0.77***

Note. *Reliability mean is based on n rater-total correlations (see column). * indicates p<0.05, ** indicates p<0.001 for SAM vs. AffectButton comparisons of reliability per scale.*

4.3 Validity

We first address predictive validity, i.e., to what extent do AffectButton measures correlate with valid *PAD* measures for the same stimuli, and concurrent validity, i.e., to what extent do AffectButton measures correlate with *PAD* values measured using SAM in the same study.

[5] T-test were performed with individual rater-total correlations as statistical units. Please note that individual rater-total correlations smaller than 0 have been removed from the analysis, as these indicate misunderstanding of the task or scale reversal. This explains the different degrees of freedom in the *t*-tests.

Please note that we report the correlation between averages of *m* stimulus ratings as rated by *n* subjects *r*(*m*-2). That is, we report correlations between averages of *P*, *A* and *D* values measured with the AffectButton and (a) valid average word ratings, (b) valid average ANET ratings, and (c) concurrent average SAM ratings. The averages in Table 3 are first per study type (not weighted as each average is based on the same number of stimuli), and then over all studies (again a simple average to not give an arbitrary high weight to a particular type of stimulus). As such, we assume each study contributes equally to the mean correlations (equal weights).

We found strong overall correlations for *P* and *D* scales, and a moderate correlation for the *A* scale (.93, .85, and .69 resp., Table 3). This indicates predictive validity and concurrent validity. Further, we see again, that the *A* scale validity is worst, especially on the situation rating study with the female sample group (ANET_SAM1). This is probably due to the lowered reliability of that scale in the ANET rating studies, although this cannot explain why the concurrent validity is higher (0.80) and comparable to the concurrent validity for the word-based studies in the same study with a gender-balanced sample (ANET_SAM2, concurrent), because in both studies the reliability was equivalent . This indicates that either the subject sample or the physical input device influenced the ratings (ANET_SAM2 was rated with a touchpad not with a mouse). Together with the lower reliability for *A*, this shows improvements should be made to the behavior of the Arousal scale, as discussed further in the reflection section. We would like to stress, though, that for the AffectButton as a whole, across studies the validity is good for *P*, *A* and *D*.

Table 3.
Correlations between average AffectButton ratings and predictive and concurrent criteria.

Study	n	m	P	A	D
Words (predictive)					
WordRating1	11	32	0.90	0.80	0.81
WordRating2	9	32	0.85	0.80	0.79
Mean [confidence]		*64*	*0.88 [0.80 – 0.92]*	*0.80 [0.69 – 0.88]*	*0.80 [0.69 – 0.88]*
ANET texts (predictive)					
ANET_SAM1	≈17	60	0.94	0.55	0.86
ANET_SAM2	≈27	60	0.95	0.63	0.87
Mean [confidence]		*120*	*0.95 [0.92 – 0.96]*	*0.59 [0.46 – 0.70]*	*0.87 [0.81 – 0.90]*
ANET SAM ratings (concurrent)					
ANET_SAM1	≈17	60	0.93	0.55	0.90
ANET_SAM2	≈27	60	0.96	0.80	0.89
Mean [confidence]		*120*	*0.95 [0.92 – 0.96]*	*0.68 (0.56 – 0.76]*	*0.90 [0.85 – 0.93]*
Mean [confidence]		***304***	***0.93 [0.91 – 0.94]***	***0.69 (0.62 – 0.75]***	***0.85 [0.81 – 0.88]***

Note. All p<0.001. Correlations 1.0<r<0.8 indicate strong correlations, 0.8<r<0.5 indicate moderate correlations (Coolican, 1990). Confidence interval for mean correlation is between brackets. The degrees of freedom for actual correlations is m-2 (see column). The basis for calculated correlation confidence intervals is m[6]*.*

We now address convergent construct validity, i.e., to what extent do measures of affect with the AffectButton relate to other theoretically related constructs (Coolican, 1990), such as preferences, musical properties and mood induction. Because the convergent constructs we tested are quite dissimilar, it is not helpful to generate one overall measure of validity as

[6] See (DeCoster, 2004) for formulas used to calculate correlation confidence intervals.

shown for the predictive and concurrent validity. Instead, we summarize results for each study (see Table 4 for a condensed view).

Table 4.
Construct validity overview.

Study	Convergent construct	Result
PreferenceRating (n=32, 9 ratings per subject, equals 288 units of measurement)	Holiday liking on 9-point Likert scale	AffectButton ratings and Likert ratings correlate, regression analysis results in significant model able to predict Likert ratings based on AffectButton *P* ratings.
MusicRating (n=21)	Objective music properties, subjective *PAD* music trace	Qualitative analysis of the traces follows feeling of songs. Correlations between values on affect dimensions and rated objective music properties are in accordance with literature. *P* and *D* behave separately.
AvatarEmotionRating (n=48)	Dominant or submissive agent's situation description	Experimental dominance manipulation confirmed with AffectButton, significant only for *D*.
Induction_SAM (n=128)	Pleasure or stressed subject mood induction	Experimental mood induction manipulation could be confirmed with the AffectButton, with the exception of anomalous behavior of the *A* scale.

In the PreferenceRating study, subjects rated holiday preferences using a 9-point Likert scale as well as the AffectButton. The AffectButton ratings and Likert ratings correlate significantly for *P* ($r(286)=0.68$), *D* ($r(286)=0.50$), and *A* ($r(286)=0.34$), all $p<0.001$. A regression analysis predicting Likert scale rating, included all three affective factors as independents, resulted in a significant model with *P* being more important than *A* and *D* (Beta's *P*, *D*, and *A*, are 0.54, 0.18 and 0.13, respectively). This is exactly as one would expect, as the expression of liking on a Likert scale is in fact an expression of positiveness versus negativeness of the preference, and as such should correlate with *P* in the first place. As shown by the regression analysis, the prediction of Likert scale ratings is significant and mainly dependent on *P*, providing convergent construct validity for *P* as measured with the AffectButton. [7]

In the AvatarEmotionRating study subjects were instructed to rate the feeling they thought the agent had, after reading a short story in which the dominance position of the agent was manipulated (powerful boss versus potential employee in need of a job). MANOVA analysis showed that subjects' *PAD* ratings of the agent's feeling were significantly influenced by the boss versus employee condition, $F(3, 44)=3.3$, $p=0.03$. Univariate analysis confirmed that *D* was significantly higher in the agent = boss condition ($M=0.28$, $SD=0.52$) versus ($M=-0.11$, $SD=0.49$), $F(1, 46)=6.9$, $p=0.01$, with *P* approaching significance ($F(1, 46)=3.9$, $p=0.054$), and *A* not being significant, $F(1,46)=0.05$, ns. This provides convergent construct validity for *D* as measured with the AffectButton because the effect of the dominance manipulation was confirmed, as well as divergent validity for *D* as a separate factor because the other two factors *A* and *P* did not significantly change due to the dominance manipulation.

[7] It is true that the Likert scale ratings also correlate with *A* and *D*, but a large portion of that correlation is corrected in the regression analysis. The remainder of the positive relation between Likert scale and *A* and *D* could (tentatively) be explained by the stimuli: one usually gets sad/disappointed about a dull vacation and enthusiastic about a nice one, not angry and satisfied respectively. Sadness and enthusiastic are *-P-A-D* and *+P+A+D* affects respectively, explaining a positive relation also on *A* and *D*.

In the MusicRating study, each participant rated each song as often as they wanted while listening to the songs. Objective music properties for each 5 seconds period had previously been scored for all three songs by one of the authors of the original publication who is a trained musician. For each song, we binned all AffectButton measurements for all participants together, and sorted them based on time stamp. Then, for each affective dimension and song we created an affect trace based on a running average over the sorted data points, resulting in three "overall" *PAD* affect traces, one for each song (see Fig. 6, Section 5.2.3). Qualitative analysis of the traces follows very well the feeling of the music (Broekens, Pronker, et al., 2010). Correlations between values on affective dimensions and rated objective music properties are in accordance with findings of others (Kellaris & Kent, 1993). These qualitative and quantitative analyses indicate convergent construct validity for the AffectButton as a whole. Further, as seen in Figure 4, there is a dissociation between the *P* and *D* factor, most notably between *t*=16 and *t*=28 where *P* goes up and *D* goes down, and between *t*=105 and *t*=115 where P goes down and D goes up. This indicates that *P* and *D* can vary independently. This again indicates divergent construct validity for *P* and *D*.

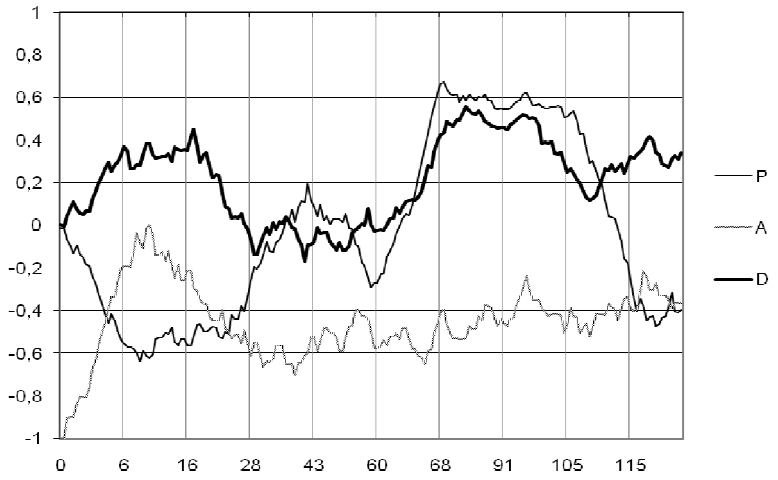

Figure 4. Affect trace of the song Walk the Plank. Note the dissociation between P and D, indicating that P and D vary independently of each other in this domain. For a detailed qualitative analysis see (Broekens, Pronker, et al., 2010).

In the Induction_SAM study, subjects underwent a mood induction that was an imagination exercise either targeted at a recent stressful event or a recent pleasurable event. MANOVA analysis showed that the induction was successfully measured with the AffectButton $F(3, 124)=79$, $p<0.001$. Univariate analyses showed that *P*, *A* and *D* were higher in the pleasure induction ($F(1,126)=511$, $p<0.001$; $F(1, 126)=52$, $p<0.001$; $F(1, 126)=78$, $p<0.001$; resp.). These effects were confirmed by the same analyses on the SAM-based ratings, with the exception of the effect on *A* that was inversed (lower *A* for pleasure induction). The induction was not specifically targeted at particular *P*, *A* and *D* effects, and only referred to the recent event to be imagined as stressful versus pleasant. Apparently this manipulated *P*, *A* and *D*. We also tested to what extent a Decision tree (QUEST method, $n=128$, 10-fold cross validation) could learn to predict if an individual *PAD* rating produced by the AffectButton belonged to a subject in the stress or pleasant induction condition. Analysis showed that

based on AffectButton ratings the decision tree predicts 85.9% correct (Binomial test $p<0.001$), while based on SAM ratings the decision tree predicts 96.1% correct (Binomial test $p<0.001$) [8]. This proved to be a significant difference in prediction accuracy (McNemar test $p<0.01$). This confirms our previous findings related to reliability and validity in a different context, i.e., that SAM is more precise.

Overall these studies show that the affective feedback gathered with the AffectButton can be used to predict values on theoretically related constructs. It seems that the only anomalous finding here is the inverted effect of pleasure versus stress induction on the *A* scale in the Induction_SAM study. Here SAM ratings showed a higher *A* level for stress induction compared to pleasure induction, while the AffectButton showed a lower *A* level for the stress induction compared to pleasure induction. We discuss this in Section 5.1.

The last types of validity we address are ecological and population validity, i.e., to what extent can we generalize our findings with respect to affect type, setting and population. As seen in Table 1, the eight studies together address 6 different types of stimuli, 3 different types of affective feedback, and 5 different types of sample groups totaling n=776. Given the fact that in each of these studies, the AffectButton performed as predicted (construct validity) or significantly correlated with other measures of affect (predictive and concurrent validity), we conclude that the AffectButton is a generic affect measurement tool, at least for Dutch natives in the age group 15 – 56. In the reflection section we discuss this aspect further, especially in light of cultural differences and universality of facial expressions.

4.4 Usability

Our analysis up until now shows that the AffectButton produces reliable and valid feedback. However, it is equally important that the AffectButton is usable. As seen in Table 5, we analyze three usability factors and aggregate across the different studies. These are *time needed to rate a stimulus* (including perception of the stimulus), the *perceived ease of use* (i.e., inverted perceived effort) using the AffectButton as rated by the participants in the study, and the *perceived liking*. Ease of use and liking averages are on a 7-point scale. We contrast the AffectButton against SAM, when this data was available in the different studies, by calculating overall corrected Cohen's effect size *d*, i.e., we report Hedges' *g*. Hedges' *g* is a measure of effect that is corrected for a bias of Cohen's *d* to produce effect sizes that are slightly too large (DeCoster, 2004). Hedges' *g* allows us to interpret overall effects of the AffectButton on effort, liking and time needed to rate across the different studies. We aggregated mean scores on liking, ease of use, time needed and *g* first per study type based on sample-size weighted average for the different statistics, then across the different studies using a simple average[9].

[8] Note that the baseline probability for correct prediction is 52.3%, i.e., the majority class predictor.

[9] Please note that since the number of studies is quite small, a random effects model for determining the weights of the individual studies gets into trouble with the between study variance (Borenstein & Hedges, 2009). However, a fixed effects model would get into trouble in our case because the studies are different, the measures are different within studies, and the number of subjects in these studies is very different. Therefore, a fixed effect model would not do justice to large effects found in small studies, and the variance term is difficult to interpret because we integrate different measures of effort within one study. Because of these complications we chose a more naïve method for integrating a summary *g*, i.e., simple averaging when studies are different (assume equal weight) and weighting based on *n* when the study is the same but only the *n* is different (between ANET studies).

As can be seen in Table 5, AffectButton usage is associated with lower actual effort as measured by the time needed to rate a stimulus (overall Hedges' *g* effect size =0.529). The average time needed is 12.9 seconds, including perception of the stimulus. Interestingly, across all studies, overall perceived ease of use is lower for the AffectButton than for other rating instruments (overall effect size *g*=0.531). Both effect sizes can be labeled as medium sizes for differences in means (Cohen, 1992). Nevertheless, the AffectButton is being perceived as equally likable as other rating instruments, none of the studies showed a preference for one or the other on average. This was confirmed in the Induction_SAM study, where the forced choice preference between SAM or AffectButton did not result in a significant preference for one or the other χ^2 (1, 128) = 1.53, p =0.22. Also in the same study, the 5-point Likert scale preference for the AffectButton versus SAM did not differ significantly from 3.0, it was 3.09, t(127)=0.80, ns. These results indicate that the AffectButton is acceptable as a rating instrument. Users have no overall preference for or against it. Users perceive the overall ease of use to be lower than SAM and Likert scale, while actual effort is also lower. It also indicates that further research is needed on why users perceive the effort to be higher. A possible explanation is that the instructions for SAM include explanation of the dimensions, whereas with the AffectButton the user has to do the interpretation in terms of dimensions. However, in the AffectButton case users are never asked for dimensions, they are asked to rate a piece of content. So, the need to interpret in terms of dimensions is not there. A better explanation is that exploration of the AffectButton's input space is needed at first time usage and this involves effort. As we know there is a strong learning effect in terms of time needed to rate stimuli (Broekens & Brinkman, 2009) in the order of 2 seconds already after 15 ratings, it could be the case that after using the button for some time, the perceived effort is much lower. This is worth investigating. This interpretation is supported by a recent study in which the AffectButton was used without explanation and possibility to practice. The study showed that the usability (rated ease of use) was significantly less (but still neutral, not worse than neutral) than the other components in the interface of which usability was tested (Brinkman et al., 2011). This indicates that the AffectButton is indeed a more complex interface component to use when unfamiliar with it. Further investigation is needed to find out the exact effect of training and instruction on the perception of effort and ease of use.

In addition to the previous quantitative analysis, we also report qualitative preferences for either the AffectButton or SAM (depending on what subjects chose in the forced choice mentioned above). The data were collected as part of the Induction_SAM study. People were asked to choose between SAM or the AffectButton. Subsequently they were asked to give *one main reason* for their choice. Results are shown in Table 6. Significant differences were found. The AffectButton is preferred for being appealing and intuitive, while SAM is preferred for being simple (nearing significance) and clear. These qualitative results provide some insight into why perceived ease of use is lower for the AffectButton than for SAM. Once the dimensions of SAM are understood, it provides a clear view of the complete input space (clear and simple), while the AffectButton needs some exploration every time affective feedback is given. Apparently individual differences in preference for affective rating instruments exist. Some prefer clear and simple ones, while others prefer fun, visually appealing and intuitive ones. This could relate to whether someone approaches rating from a cognitive or an affective point of view. This is an important area of further study.

Table 5.
Usability factors

Study	n	SAM	AffectButton	t	p	Effect size g
Mean time per rating (sec)						
Words						
WordRating1	11		10.9			
WordRating2	9		11.6			
Weighted average Words	*20*		*11.2*			
Music						
MusicRating	21		13.5			
ANET texts						
ANET_SAM1	202	16.0	12.0	4.93	<0.01	0.693
ANET_SAM2	325	17.5	15.0	3.85	<0.01	0.427
Weighted average ANET	*527*	*16.9*	*13.9*			*0.529*
Average of Words, Music & ANET		***16.9***	***12.9***			***0.529***
Perceived ease of use						
ANET texts						
ANET_SAM1	202					
Ease of use		5.44	4.85	2.77	<0.05	0.390
Perceived effort (inversed)		5.21	4.50	3.25	<0.05	0.457
Average ANET_SAM1		*5.33*	*4.68*			*0.424*
ANET_SAM2	325					
Ease of use		4.18	3.62	3.26	<0.05	0.362
Perceived effort (inversed)		4.48	4.71	-1.20	0.23(ns)	-0.133
Average ANET_SAM2		*4.33*	*4.17*			*0.114*
Weighted average ANET	*527*	*4.71*	*4.37*			*0.233*
PreferenceRating		Likert				
Perceived effort (inversed)	32	5.25	4.09	4.69 (paired)	<0.001	0.829
Average of ANET & Preferences		***4.98***	***4.23***			***0.531***
Perceived liking						
ANET texts						
ANET_SAM1	202					
Pleasant to use		5.42	5.13		ns	
Easy to understand		5.60	5.52		ns	
Average ANET_SAM1		5.51	5.33			
ANET_SAM2	325					
Pleasant to use		3.75	3.80		ns	
Easy to understand		3.78	3.97		ns	
Average ANET_SAM2		*3.77*	*3.89*			
Weighted average ANET	*527*	*4.44*	*4.44*			
PreferenceRating		Likert				
Liking	32	3.94	4.19	(paired)	ns	
Average of ANET & Preferences		***4.19***	***4.32***			***---***

Table 6.
Qualitative differences found for reported reason for forced preference (count)

Reason	***AffectButton***	***SAM***	χ^2	***p***
Precise	10	13	0.39	0.53
Simple	9	19	3.57	0.059
Clear	9	28	10.4	0.001
Fun/visually or affectively appealing	18	1	15.2	<0.001
Intuitive (recognizable/personal/direct)	5	0	5.00	0.025
Reliable/valid/representative	1	3	1.00	0.32
Other	5	7	0.33	0.56
Total (128)	57	71	1.53	0.22

4.5 Relations between Pleasure, Arousal and Dominance

In addition to reliability, validity and usability results, we report findings on the relation between *P*, *A* and *D* that support the design rationale of the AffectButton. We found that very positive or negative affective experiences tend to have a high *A* component, a finding compatible with findings of others, e.g., (Britton et al., 2006). This is illustrated in Figure 5 (right) showing a *P-A* scatter plot of a pooled set of data points from valid stimuli from the databases ANET, ANEW, IAPS, IADS (Bradley & Lang, 2007; Lang, 2008), and Mehrabian's word list (Mehrabian, 1980), and SAM-collected data in our studies Induction_SAM, AvatarEmotionRating and ANET_SAM1 (n=480, 60 from each data set). The reference lines show the *P-A* relation as implemented in the AffectButton projected onto the pooled data. Data points left and right of the reference lines are sparse. This indicates that there is indeed a lack of very high and very low valence measures in combination with low arousal. This finding replicates and extends earlier findings investigating the two dimensional P-A shape (Bradley & Lang, 2000) by showing that this relation holds for a diverse set of rated stimuli. This finding supports our decision to put Arousal control in the inner border, not *D* or *P*, i.e., extreme values for *P* or *D* go together with high values for *A*.

Pleasure, Arousal and Dominance are not independent in practice. In the same pooled dataset as described above, we found *P* and *A* to correlate with r=-0.31, *P* and *D* to correlate with r=0.71, and *A* and *D* to correlate with r=-0.26 (all p<0.001). The relative absence of low *A* values coupled to very negative or positive *P* values indicates a non-linear relation between *P* and *A* favoring higher levels of arousal in combination with higher levels of pleasure. This was confirmed by a quadratic regression explaining a significant proportion of variance, R^2=.22, $F(2, 477)$=67.8, p<0.001. The moderate correlation between *P* and *D* suggests there is a linear relation between *P* and *D*. This means in any case that *P*, *A* and *D*, when used in practice, cannot be considered orthogonal dimensions. However, this does not mean they are not meaningful: *absence of orthogonally is not equivalent to absence of meaning*. There are many situations (and our MusicRating study is only one of them) in which *P* and *D* are dissociated, i.e., high *D* and low *P*, and vice versa. This indicates that *D*, as a separate factor from *A* and *P*, is a meaningful factor to describe the affective content of a situation, feeling, or stimulus (Broekens, 2012).

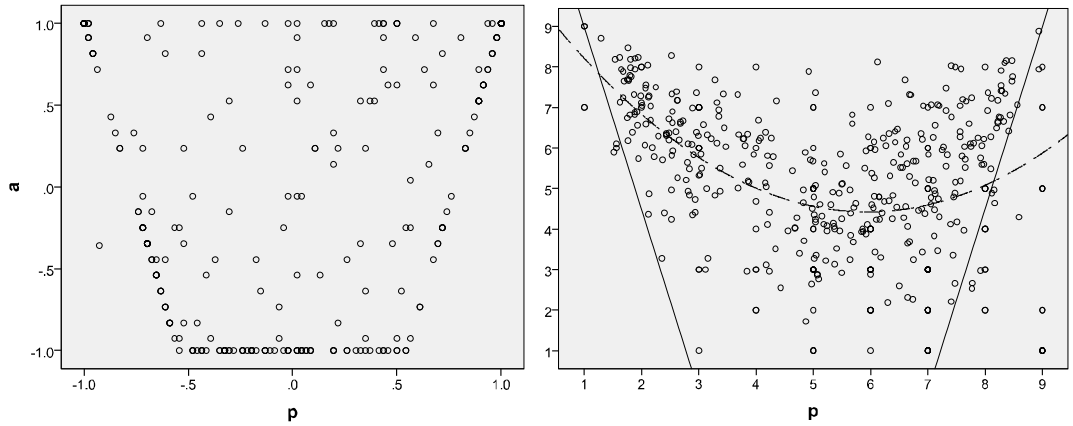

Figure 5. Pleasure-Arousal scatter plot of WordRating (left), and, (right) pooled various sources (see text). Reference lines show the P-A relation as implemented in the AffectButton projected onto the pooled data. Curve shows a quadratic fit $A=9.9-1.8P+1.6P^2$.

5. Discussion of results, reflections and lessons learned, and impact

We now present the summary of the results as well as a more detailed discussion of the rationale behind, limits of, and lessons learned while developing the AffectButton. Then we discuss the potential impact of the AffectButton for affective HCI. Throughout this discussion opportunities for future work are discussed.

5.1 Summary of the results

The main conclusion that can be drawn from the analysis is that the AffectButton is a reliable, valid and usable means for gathering affective self-report. This is indicated by the following:

- Measures on *PAD* factors are *reliable* as indicated by rater-total correlations (0.81, 0.66, and 0.77 for *P*, *A*, and *D* respectively).
- Measures on *PAD* factors are *valid* as indicated by (a) overall correlations between AffectButton measurements and SAM measurements as well validated stimuli (0.93, 0.69 and 0.85 for *P*, *A* and *D* respectively), (b) consistent predictive power of AffectButton ratings for related constructs such as preferences, mood manipulation, and music, (c) a large sample size of 776 subjects of varying demographic nature showed to be able to use it for a variety of types of stimuli and multiple types of affective feedback.
- *Usability* figures show no significant differences between preference for the AffectButton and preference for other rating mechanisms, and lower actual effort than SAM in terms of time needed to rate a stimulus (overall effect size g=0.529).
- The amount of screen space used by the AffectButton is less than for SAM and other affective self-report methods; rating with the AffectButton is quicker then with SAM, and achieved with less instruction.

However, our analysis also indicates several points of attention. First, the reliability of the *A* scale indicates the need for further study. Another indication for the need to further study the *A* scale is that we found an inverse effect of stress versus pleasure induction when compared to SAM *A* scale ratings in the Induction_SAM study. We hypothesize this is due to a combination of two things. First, a negative induction can cause feelings of sadness and frustration, both low *A* states, and feelings of anger and nervousness, both high *A* states. The faces of the AffectButton that best relate to sadness and frustration have low *A* values. So, when a subject imagines a stressful situation (such as a conflict at work) and feels sad about it, the AffectButton will measure the sadness reaction resulting in a low *A* value. However, if the induction is rated per factor as is done with SAM, the process of rating is different and based on a cognitive decision of how much a particular situation invokes arousal. As there is no representation of prototypical emotions in SAM, and as stress is commonly known to increase arousal and subjects are explained the SAM rating scales, subjects should indeed report higher arousal levels for stressful situations. It seems that for a negative induction that is not explicitly controlled for arousal effects, the AffectButton has a bias towards lower Arousal values, while SAM has a bias towards higher values. This is the result of the rating process. Second, as subjects rated the positive induction as very happy with the AffectButton (mean P=0.60, mean D=0.63), the *A* measures were also high. This combination leads to the aforementioned inverse effect on the arousal scale. Analysis showed that the negative induction *did* increase the amount of arousal when compared to the subjects' neutral state as measured just before the induction ($F(1,66)=8.1$, $p<0.01$) as would be expected. This

confirms our idea that our finding of a reversed effect on arousal can be explained by the positive induction being rated as very happy. This indicates that either the induction was not specific enough, or that the AffectButton's happy faces with high arousal should be made to look more extreme to avoid having people selecting a happy face with high arousal unwanted. Further study is needed.

Second, the subjective ease of use is lower for the AffectButton than for other rating mechanisms. This should also be investigated, especially whether this is a novelty effect or not. As the AffectButton is a new interface component introducing a new affordance in the form of a dynamic reaction to mouse movement *inside a button*, users might need to get accustomed to it. The effect of training and/or explanation before using the AffectButton must be studied in more detail. We have shown that repeated usages of the AffectButton results in less time needed to use it (Broekens & Brinkman, 2009), but we do not know, for example, if practice influences perceived ease of use or maybe even reliability and validity.

Third, users seem to prefer SAM for different reasons than the AffectButton, which might be a good starting point for further studying the relation between interpersonal differences and the reliability, validity and usability of the AffectButton and SAM.

5.2 Reflections

We now continue with a detailed discussion of the rationale and limitations of the AffectButton. We start by framing our results in a discussion on pictorial facial expression of affect related to culture and universality of emotion expression as well as age.

5.2.1 Practical and theoretical issues and limitations.

The fact that the AffectButton presents one facial expression not only has benefits, but can also have drawbacks. We discuss four practical issues, after which we discuss the potential impact of (non-)universality of facial expressions on the AffectButton's applicability. First, if a visual stimulus or situation is presented that includes a facial expression the user might be tempted to try to match the face in the stimulus, not the affective content of the stimulus. In many cases, this should not be a problem, but in cases where the expression contradicts the affective content, it can. For example, if someone cries for joy, it might be tempting to rate the situation as sad, because the person cries. Further research is needed in this area. Second, users might be tempted to rate too extreme, because they try to match their affect to the prototypical emotion. This has an impact on reliability, because the variation in the output data is much more if users tend to rate only with extreme values of *PAD*. This has two related problems: (a) it is often not possible to find a prototype that matches affect, for example, jealousy cannot be matched to a prototypical emotion, but should be matched to a face that best approximates the feeling of jealousy instead; (b) when it is possible, the *PAD* values are to extreme, for example, happy being rated as elated. The influence of the AffectButton's facial expression must be studied in more detail, especially with respect to interpersonal differences in interpretation. Third, users might not take enough time to explore the button's input range. As a result, they will not be able to find an appropriate face that corresponds best to their feeling. This will impact validity. For example, if a user feels positively excited, but only explores the top part of the button, (s)he will end up with a strongly dominant version of excitement (elated), even if (s)he felt in awe (low dominance). These three issues might be solved with proper instruction, but instruction is exactly one of

things we wanted to not depend on too much. Up until now all users were at least prompted to play around with the button before starting the experiment. We are happy with the results obtained given this limited usage instruction. However, we do not know what the effect of instruction is in general on the validity, reliability and usability of the AffectButton, and this should be further explored. Fourth, we do not know the impact on validity, reliability and usability of changing the button's appearance, for example by scaling the AffectButton to another size than 100x100 pixels, or by adapting its design so that it fits within the style of the interface (imagine using it in a computer game or web shop and changing the looks of the face). This last issue points to the need to experimentally address computational embedding, i.e., can the AffectButton indeed be easily embedded in a wide variety of interfaces, allowing small changes that do not impact reliability, validity and usability. Computational embedding is about being able to fit the tool in an interface, as well as about the tool's ability to produce data that is meaningful and easy to interpret automatically. We have not experimentally addressed computational embedding in this paper, in the sense that we methodically varied interfaces that use the AffectButton. We consider the evaluation of this criterion to be future work. However, the AffectButton's design has been focused around solving this challenge.

An important overall limitation is that maybe affect shouldn't be measured this way. Affect might need to be measured on separate scales that aren't biased towards the extreme prototypes. Currently we have to conclude based on the large set of different studies presented, that we *can* measure affect in this way. However, and as indicated, there could be cases in which using the AffectButton is problematic and separate scales are a better alternative. In fact, we do not know if the AffectButton is biased towards extreme prototypes. In some cases it can be, but up until now we did not find evidence for that. Our remark about prototype bias is about a matching bias: a user clicks the AffectButton based on a facial match instead of a match based on affect present in that face. This potential misuse of the AffectButton must be investigated, as indicated previously. However, such biases are present in all self-report instruments. For example, the iconic representation of *dominant* in SAM is a large puppet. Large people on pictures would probably be rated as dominant, while there are certainly cases where this would introduce an unwanted bias. Despite this potential bias SAM seems to work quite well.

The question of whether the recognition and interpretation of facial expressions is dependent on culture and age is an important one for the AffectButton, although from a slightly different perspective. The classical debate is about whether humans can cross-culturally recognize (i.e., label) spontaneous facial expression of emotion. In contrast, the AffectButton aims at a consistent attribution of affect (in our case *PAD* values) to a set of carefully constructed (dynamically changing) faces, not the attribution of categorical emotions to spontaneous emotional expressions. It is true that the AffectButton uses a pictorial facial representation of affect, but we took this modality as basis for our stimuli (and in essence the faces are nothing more than *PAD*-labeled stimuli) for the following two reasons. First, posed facial expressions are a powerful means for expressing affect (Elfenbein & Ambady, 2002), regardless of whether spontaneous emotion expressions are universally recognizable. Second, abstracting away to the iconic helps recognition because iconic representations abstract away non-essential elements thereby emphasizing more essential ones (see e.g., (Desmet, 2005) for a similar reasoning). The value of this iconic facial

approach is also exemplified by the many rating mechanisms that are based either completely or partly on facial expression of affect (see review section). We think the universality issue *per se* is not relevant for the AffectButton. However, age and cultural differences are important for the generality of our claim that the AffectButton is in principle widely usable.

With regards to culture, research (Scherer & Wallbott, 1994) has shown that the effect of emotion (7 emotions) on motor expression variation versus the effect of country (37 countries) was in the order of 3-4 times higher. Even though this study did not focus on facial expression of emotion, it shows that when it comes to motor expression of emotion, a considerably larger proportion in variation is explained by the emotion than by country. This is a strong indication that emotion expression is indeed universal to some extent. Later, a meta analysis (Elfenbein & Ambady, 2002) confirmed that facial expressions presented as pictures but also as videos of faces, which comes perhaps closest to a dynamically changing iconic facial expression, can to a large extent be cross-culturally recognized, see Table 9 in (Elfenbein & Ambady, 2002). In general, imitated facial expressions (the expresser imitates a picture or muscle movement that the recognizer must then recognize) are better recognized than posed or spontaneous expressions (Elfenbein & Ambady, 2002). This finding supports our approach, as the AffectButton's expressions are iconic and specifically constructed to enhance recognition. Finally, the same review concluded that the closer cultures are, the smaller the in-group advantage (i.e., the gain individuals have in recognition accuracy of expressions expressed by people within their culture versus people outside of their culture). In addition to this, a systematic review of cross-cultural emotion recognition studies (Russell, 1994) showed that a potential problem might not be that it is difficult to recognize affect in the facial expression of a person from another culture, but to attribute the emotion label to it. More generally speaking, the focus on emotion words and categories often used in the stimulus categorization, selection and recognition rating might be problematic. An alternative account for "universal affect recognition" is proposed in (Russell, 1994), i.e., that underlying dimensions can be recognized in facial expressions. It is known that humans consistently attribute dimensional affect to a wide variety of stimuli. This is exemplified by the different stimulus sets for affective sound, words, pictures and texts developed by Bradley and Lang over the years, but also the vast amount of work on affective rating of music, environments and objects by Mehrabian and Russell. It is within this theoretical framework that the AffectButton must be seen. Whether affect is consistently attributed to subtle variations in facial emotional expression is in fact an open question, but the positive results obtained with the AffectButton are in essence a large scale confirmation of this proposal, showing that this is indeed plausible. Together, this indicates that our current sample of Dutch subjects should not pose a problem with respect to generalization of our results to close cultures, i.e., Western cultures, although this claim should be tested in future work.

With regards to age we can be more precise. Our studies show that a wide variety of age groups can use the AffectButton producing accurate measurements. Our combined subject sample was aged 15 (teenagers average minus standard deviation) to 56 (Dutch women average plus standard deviation) years covering everything except children and elderly. This means that we currently cannot and do not generalize our findings to elderly and children, and this issue remains to be researched. A problem here is that the development of

interactive products for children requires adapted design and analysis methods (Markopoulos et al., 2008), and as such the standard tests and comparisons cannot be performed when analyzing reliability and validity of the AffectButton (such as comparing SAM ratings with AffectButton ratings or asking children to rate emotion words). A potentially interesting route would be to use measurement instruments for emotion specifically developed for young children, such as Sorémo (Girard & Johnson, 2009) or elements of the Fun Toolkit such the Smileyometer (Read, 2008), and compare ratings of children on both rating instruments.

5.2.2 Measuring Pleasure Arousal Dominance factors.

Affect is not equal to mood, emotion or attitude, but any mood, appraised stimulus or emotion, i.e., any affective experience, has an approximate *P*, *A* and *D* value. The inverse is not true: many affective experiences share the same *P*, *A* and *D* values. A limitation of the design of the AffectButton is thus that, for example, emotions that are mainly differentiated by their cognitive counterpart, such as jealousy and cold anger, cannot be uniquely selected in *PAD* space, because their affective part is the same. A second limitation of the design of the AffectButton is that across affect type, values can also be the same. This means it is impossible to interpret a [-1, -1, -1] *PAD* triplet as sadness, unless you know what the framing for the rating was, i.e., you know the subject was asked to rate his/her emotion (and not, e.g., a preference for a holiday). However, these limitations belong to the underlying factor-based approach, and the same holds for all affect measurement techniques that use factors. As such we consider the AffectButton a method for affective self-report, not emotion self-report per se.

5.2.3 Arousal scale behavior and alternative 2D to *PAD* mappings

As shown in the results section, the reliability of the *A* scale as measured with the AffectButton is acceptable but not very good. Further, individual (not stimulus average) Arousal ratings have a rather unnatural distribution (Fig. 6), caused by the fact that Arousal is the derived dimension and only manipulated at the edges of the button. Apparently, with the current button design users tend to rate Arousal at either negative or positive extremes. As a result, the central point of the Arousal distribution is not equal to 0.0. This is also exemplified by the Arousal trace in Table 4 (MusicRating experiment). The trace shows that users rated the music as calm, even though subjectively the first part of that song would definitely merit a positive arousal rating. So, there seems to be a distribution and reliability problem with the Arousal scale, probably due to it being the "derived" dimension in the AffectButton, while Pleasure and Dominance are the "primary" dimensions. In an attempt to tackle both problems (distribution and reliability), we tried three alternative 2D to *PAD* mappings. The first one (in fact tried before the 2D mapping) is a real 3D mapping with *A* on the scroll wheel. In a preliminary study this showed not to be usable, as users simply did not associate a button with scroll-wheel affordance (Broekens & Brinkman, 2009). Further, this solution does not scale to touch pads, touch screens, video game controllers, and other devices that lack a third input dimension. As such, we dropped this entire idea. A second alternative was based on the relation between *P*, *A* and *D* as empirically found in the previously mentioned pooled data set (see previous section). We performed a factor analysis on the three factors to extract 2 components (that were used as *x* and *y*). The extracted components explained 91% of the variance, and we used the factor loadings to control *P*, *A* and *D*. The third alternative mapping is based on a circumplex layout, where *P* and *D* are still

controlled by X and Y respectively, but *A* is controlled by a sine function (taking the radial based on *PD* as input) multiplied by the distance from (0,0). This results in a wave alternating between high and low *A*, with higher amplitude (more extreme values of *A*) at the edges of the AffectButton. The function was such that in each *PD* quadrant the wave has a complete period, so that in each *PD* quadrant it is possible to find both high and low *A* values. Both of the mappings resulted in lower reliability for *P*, *A* and *D*, based on initial testing on the same stimuli as those used in the WordRating1 and 2 studies. As such, both were dropped for now. We conclude that the *A* scale is functional, but not optimal, and that further research is needed into the tradeoff between usability, reliability and intuitiveness of the resulting data. Especially this last point is important, as currently the *A* scale may produce individual measurements that are not intuitive (very high or very low), even though the resulting average values are valid.

An important problem with (small) changes in an instrument's workings is that to rapidly assess the effect of different versions on the reliability and validity of the measurements, a standard benchmark set of affective stimuli for testing is needed containing a balanced but limited set of stimuli. This is needed to rapidly assess the effects without redoing a whole batch of experiments. There is no such benchmark at this point, and we are working on one that is based on a subset of the previously mentioned pooled dataset of different sources with the possible addition of video samples.

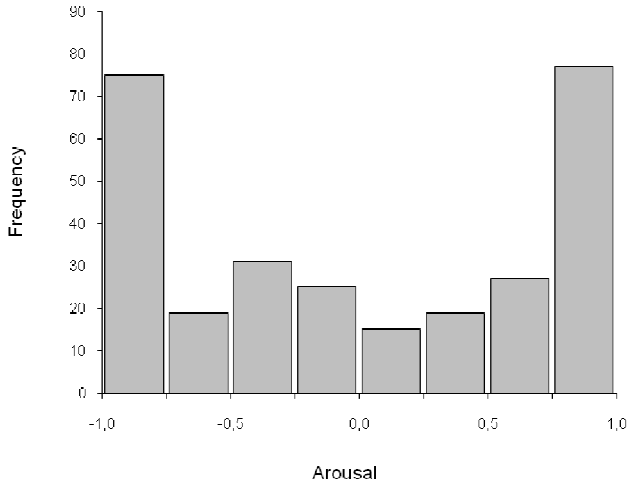

Figure 6. Histogram of individual Arousal values in WordRating2

5.2.4 AffectButton is not equal to SAM

Although it is tempting to conclude from our analysis that the AffectButton is intended as a replacement for SAM, this is not the case. SAM and the AffectButton are different measurement tools, aiming to measure the same construct. SAM has benefits and drawbacks, just as the AffectButton. Several key differences include, first, that the AffectButton is state based, meaning that one sees the currently selected affective state, but not the *PAD* space, as in SAM. Second, the AffectButton shows a rendered face, and as such does not need understanding of affective dimensions and cognitive reflection upon the meaning of an affective experience in terms of *P*, *A* and *D*. Third, the AffectButton is a small

device, needing considerable interaction possibilities, while SAM needs a considerable amount of space but can also be administered on paper. Fourth, SAM is more reliable than the AffectButton, indicating that when ease of use, usage by groups of people unable to cognitively reflect upon their affective state, or embedding in an interface is not needed, SAM is a better alternative. Fifth, we found individual differences with respect to preferring SAM over the AffectButton. These tradeoffs indicate it is useful to frame the selection of an affective self-report method in a task and population context.

5.2.5 Mixed emotions

In the introduction we mentioned that a factor-based approach can handle mixed emotions better than a categorical approach. This is only partly true, and we have to be explicit about what we mean by mixing. There are two possible meanings relevant to the current discussion. Mixing as in integrating different measurements, and mixed emotions as in feeling happy about something and sad about something else at the same time. Factor-based representations, such as used for the AffectButton and SAM, are a common representational format of affect that can be used to integrate different emotions in a sensible way, by using aggregation functions such as averaging. This results in a meaningful representation of the affective state that can subsequently be reasoned upon or calculated with again. However, while doing so, the mix "consumes" the individual measurements. For example mixing happy and sad would result in the average of both measurements unless one retains the originals. In a categorical approach the originals are always retained as integrating measurements of different emotions cannot be done. Mixing as in feeling two different emotions at the same time (or quickly alternating) can in the case of the AffectButton only be coped with if the person is asked to rate both emotions, just like he/she would have to do when using a measurement device based on categories. With the AffectButton, a person would select for each category of the mix a different expression. We do not know how it performs in this case. Measuring mixed emotions is an interesting topic for further topic.

5.2.6 Measuring mood, attitude and emotion

A notable exception to what our studies cover is the measurement of actual emotion, i.e., the short lived, intense version of affect. We do measure mood in the last study, and attributed affect in the other studies, but not the short-lived intense form of affect referred to as emotion, such as when someone is very angry. We do not know if the AffectButton is usable in that setting. However, this is a shortcoming shared with almost all self-report instruments of emotion. Many articles talk about measuring emotion, while in fact mood, attitude about a stimulus, or affect present in a stimulus is measured. For example, any induction study is about mood, not about emotion, any media content rating study is about rating the combined affective effect of the stimulus on the user, and any preference elicitation study is about attitude, not emotion. We have not yet done so, like many others have not, as this is very difficult to do. For example, it is difficult to get someone in an actual emotional state in a controlled setting while adhering to ethical guidelines. Also, asking a subject to rate its emotional state, but not the affect contained in the eliciting stimulus, is methodologically thorny. Further, there is an enormous amount of usage potential, even when the AffectButton is used "only" for the measurement of mood, preferences and affect attribution as this is what is typically needed in usage scenarios related to rating and annotation (music/movie preferences, product evaluations, written situations, affect priming

studies, etc…). As such we consider this not so much a limitation for the AffectButton but more so a challenge for future research in affective self-report methods.

5.3 Discussion of impact

Because of the AffectButton's simple and compact design, and because of it being an open tool (http://www.joostbroekens.com) of which the workings are transparent, it can be used for a wide variety of applications, and even could be ported to other devices (although additional validation is needed in that case). We feel it is opportune to discuss the AffectButton's possible usage, and hence, potential impact.

5.3.1 Applications where user affect is needed for computation and communication

If user affect is needed as input for a function of a computational system, such as for user modeling, the AffectButton is well usable. It is compact and therefore easy to embed in an interface in its current form. Further, the data it produces is computation friendly. Affect factors are meaningful separately as well as combined, and measurements can easily be integrated and compared over time, and across users (e.g., for user mood, user group mood, affective trends over time, mood clusters, intra- and interpersonal differences). In general, whenever affective preferences need to be collected, and one chooses for a factor based approach, the AffectButton can be used. Recent studies show that affective preferences need to be considered (Pommeranz et al., 2012), and indeed can be collected with the AffectButton (Broekens, Pommeranz, et al., 2010; Pommeranz et al., 2008).

Affective preferences are relevant for a wide variety of applications. For example, in the field of recommender systems (Burke, 2002; Herlocker et al., 2004), the average mood associated with a song could be calculated based on AffectButton data. Songs can then be recommended to people based on mood data, as seen in (Janssen et al., 2012). User affect is also relevant for social peer-to-peer systems (Pouwelse et al., 2008). Related to recommendation is content annotation. An important reason for users to annotate is to express themselves to their peers (Ames & Naaman, 2007), and emotion expression is an important element in computer mediated communication in general (Derks et al., 2008), as witnessed by the spontaneous emergence and use of emoticons (Walther & D'Addario, 2001). The AffectButton can be used to fill in the need to communicate affectively, while at the same time it provides a means for valid and reliable feedback to the system. For example, in affective video annotation (Chen et al., 2008; Soleymani & Larson, 2010), the AffectButton could be used for affect self report in a relatively unobtrusive way (analogous to our MusicRating experiment).

5.3.2 Applications where user affect is needed for assessment

The AffectButton can also be used in applications involving user assessment. This includes public opinion polls, online psychological coaching, affective tutoring agents, customer and employee satisfaction polls, and affective trend analysis. For example, the AffectButton can be used as an alternative to question-based affective self-report or automatic recognition for the extraction of the student's affective state in a tutoring system that tracks and adapts its functioning to the user's affective state (D'Mello et al., 2007; D'Mello et al., 2008; Woolf et al., 2010). Common learner emotions such as confidence, excitement, boredom and anxiety are represented in distinct areas in *PAD* space, so if one is interested in affect factors we think the AffectButton is suitable as a measurement tool in this context. For opinion polling,

the AffectButton provides an alternative to affect questionnaires and produces easy to integrate units of measurement (*PAD* triples) that can be assessed per scale, but also aggregated together and assessed on the three scales together. It is particularly well suited for this type of application because of the need for data aggregation on an abstract level facilitated by the factor-based approach. An important related issue is that after *PAD* data collection, this data might need to be visualized, such as for online coaching and treatment (Brinkman et al., 2010): the therapist needs to react upon the client's affective feedback. A possible way could be to use the AffectButton again, but now as output of (aggregated) data. This issue is interesting future research related to data visualization.

5.3.3 Relevance for experimental psychology and psychometrics

As mentioned in the introduction, studies in experimental psychology and psychometrics traditionally have an interest in self report of emotion and affect. For example studies that are based on the induction of affect, such as studies investigating the effect of affect on creativity (Isen et al., 1987) on negotiation (Anderson & Thompson, 2004; Carnevale & Isen, 1986; George et al., 1998; Kleef et al., 2006) on cognitive control (Dreisbach & Goschke, 2004; van Steenbergen et al., 2010) and information processing mode (Gasper & Clore, 2002) always have the need to do an affect manipulation check. This means that subjects need to be questioned in order to verify if the affective manipulation had any effect, prior to investigating the hypothesized effect of affect on cognitive control, creativity, etc. The AffectButton can be used as a tool to easily measure affect in this context. Care has to be taken when interpreting relations between arousal and the other two dimensions in this case, as arousal in the AffectButton is dependent on Dominance and Pleasure. However, given the recent revival of the dominance (potency) dimension as an important one in affect studies and the reliable and valid results obtained with the AffectButton on measuring Dominance and Pleasure, we feel the AffectButton is a timely and welcome addition to the self-report instruments that are already available and often used (Bradley & Lang, 1994; Russell et al., 1989; Watson et al., 1988).

5.3.4 Using the AffectButton with other input devices

Our research shows that the AffectButton can be used with a mouse and a touchpad. Other devices to control its behavior could include game controllers and touch screens. There are two main requirements for such devices. First, the AffectButton requires input in two dimensions. Second, the precision must be at about 7 bits per axis, resulting in 128 possible values per axe which is roughly equivalent to the evaluated size of the AffectButton. This means the AffectButton can in principle be controlled with a wide variety of physical input devices including a mouse, a touchpad, a touch screen (directly), or the analog stick of a game controller. Especially control with an analog stick can be interesting. First, such sticks give a bit of force feedback giving the user perhaps even more intuition about extreme emotions being at the edge of the button. Second, the possibility of controlling the AffectButton with an analogue stick opens the way towards affective feedback on game consoles that is reliable and valid. With the current broadening of functionality of game consoles including movie playing, picture viewing, browsing the internet, and social networks, the ability to integrate affective feedback becomes even more interesting and relevant. However, the influence of the input device on reliability, validity and usability is not known and should be studied.

6. Conclusion

In this article we have analyzed 8 studies with the AffectButton with a total of 776 subjects. Our analysis shows that the AffectButton is a reliable, valid and usable method for affective self-report. The AffectButton should be seen as an alternative to the Self-Assessment-Manikin, with the AffectButton focusing on usability and computational and interface embedding enabling 1-click affective self-report on 3 dimensions. Further the AffectButton has Pleasure and Dominance as primary dimensions, and Arousal as "derived" dimension. This makes it particularly useful when dominance-related feedback is needed because there is no need to explain the dominance dimension to users. It is not a replacement of SAM in all possible usage scenarios, as argued in the discussion. We have also given a detailed explanation of its theoretical origins and limitations. Among these limitations we want to stress the following: (1) the AffectButton has been tested with Dutch subjects only, (2) because of its design the AffectButton cannot be used to study correlations between Pleasure and Arousal or Dominance and Arousal, (3) we tested the AffectButton for a wide variety of experienced affect (mood, attitude, preference) but not for experienced affect resulting from the subject's own emotion, and (4) the AffectButton has been developed for gathering *affective* self report, not for the self report of emotion categories. Finally, we have extensively discussed future research including the AffectButton's usage potential in a wide variety of applications related to preference elicitation and user assessment, as well as the possibility of using other input devices than mouse or touch pad for controlling it.

Acknowledgments

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 VICI-project 08075. It has further been supported by the Dutch broadcasting companies VPRO and NTR, as part of the "Groot Nationaal Onderzoek". We thank Wietske Visser, Gerwin de Haan, Robin Read and Sylvie Girard for their constructive feedback and discussion. Further, we thank the following persons for carrying out the work for several of the experiments: Inge Knippenberg, Janneke van der Zwaan, Alina Pommeranz, Eddie van der Wereld, Wim van der Ham, Anne Pronker and Marjan Neuteboom.

References

Ames, M., & Naaman, M. (2007). Why we tag: motivations for annotation in mobile and online media *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 971-980). San Jose, California, USA: ACM.

Anderson, C., & Thompson, L. L. (2004). Affect from the top down: How powerful individuals' positive affect shapes negotiations. *Organizational Behavior and Human Decision Processes, 95*(2), 125-139.

Bartneck, C., Reichenbach, J., & Breemen, A. (2004). In your face, robot! The influence of a character's embodiment on how users perceive its emotional expressions *Proceedings of Design and Emotion 2004* (pp. 32–51).

Baumeister, R. F., Vohs, K. D., & Nathan DeWall, C. (2007). How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation. *Personality and Social Psychology Review, 11*(2), 167.

Bickmore, T. W., & Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Trans. Comput.-Hum. Interact., 12*(2), 293-327.

Borenstein, M., & Hedges, L. V. (2009). *Introduction to meta-analysis*: Wiley.

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the Self-Assessment Manikin and the Semantic Differential. *Journal of Behav Ther Exp Psychiatry, 25*, 49-59.
Bradley, M. M., & Lang, P. J. (2000). Affective reactions to acoustic stimuli. *Psychophysiology, 37*(2), 204-215.
Bradley, M. M., & Lang, P. J. (2007). Affective Norms for English Text (ANET): Affective ratings of text and instruction manual. *Techical Report. D-1, University of Florida, Gainesville, FL*.
Brave, S., Nass, C., & Hutchinson, K. (2005). Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International Journal of Human-Computer Studies, 62*(2), 161-178.
Breazeal, C. (2003). Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies, 59*(1-2), 119-155.
Brinkman, W. P., Hattangadi, N., Meziane, Z., & Pul, P. (2011). Design and Evaluation of a Virtual Environment for the Treatment of Anger *Proceedings of Virtual Reality International Conference (VRIC 2011)*.
Brinkman, W. P., Mast, C., Sandino, G., Gunawan, L. T., & Emmelkamp, P. M. G. (2010). The therapist user interface of a virtual reality exposure therapy system in the treatment of fear of flying. *Interacting with Computers, 22*(4), 299-310.
Britton, J. C., Taylor, S. F., Sudheimer, K. D., & Liberzon, I. (2006). Facial expressions and complex IAPS pictures: Common and differential networks. *NeuroImage, 31*(2), 906-919.
Broekens, J. (2007). Emotion and Reinforcement: Affective Facial Expressions Facilitate Robot Learning *Artifical Intelligence for Human Computing* (pp. 113-132).
Broekens, J. (2010). Modelling the experience of emotion. *International Journal of Synthetic Emotions, 1*(1), 1-17.
Broekens, J. (2012). In Defense of Dominance: PAD Usage in Computational Representations of Affect. *International Journal of Synthetic Emotions, 3*(1), 33-42.
Broekens, J., & Brinkman, W.-P. (2009). AffectButton: Towards a Standard for Dynamic Affective User Feedback *ACII 2009*: IEEE.
Broekens, J., Pommeranz, A., Wiggers, P., & Jonker, C. M. (2010). Factors Influencing User Motivation for Giving Online Preference Feedback *5th Multidisciplinary Workshop on Advances in Preference Handling (MPREF'10)*.
Broekens, J., Pronker, A., & Neuteboom, M. (2010). Real time labeling of affect in music using the affectbutton *ACM Workshop on Affective Interaction in Natural Environments* (pp. 21-26): ACM.
Brown, J. H., & Maurer, B. A. (1986). Body size, ecological dominance and Cope's rule. *Nature, 324*(6094), 248-250.
Burke, R. (2002). Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction, 12*(4), 331-370.
Calvo, R. A., & D'Mello, S. (2010). Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. *Affective Computing, IEEE Transactions on, 1*(1), 18-37.
Carnevale, P. J. D., & Isen, A. M. (1986). The influence of positive affect and visual access on the discovery of integrative solutions in bilateral negotiation. *Organizational Behavior and Human Decision Processes, 37*(1), 1-13.
Cashdan, E. (1998). Smiles, speech, and body posture: How women and men display sociometric status and power. *Journal of Nonverbal Behavior, 22*(4), 209-228.
Chen, L., Chen, G.-C., Xu, C.-Z., March, J., & Benford, S. (2008). EmoPlayer: A media player for video clips with affective annotations. *Interacting with Computers, 20*(1), 17-28.

Cohen, J. (1992). A Power Primer. *Psychological bulletin, 112*, 155-159.
Conati, C., Marsella, S., & Paiva, A. (2005). Affective interactions: the computer in the affective loop *Proceedings of the 10th international conference on Intelligent user interfaces* (pp. 7-7). San Diego, California, USA: ACM.
Coolican, H. (1990). *Research methods and statistics in psychology*: Hodder & Stoughton Educational.
Core, M., Traum, D., Lane, H. C., Swartout, W., Gratch, J., van Lent, M., et al. (2006). Teaching Negotiation Skills through Practice and Reflection with Virtual Humans. *SIMULATION, 82*(11), 685-701.
Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schroeder, M. (2000). FEELTRACE: An instrument for recording perceived emotion in real time *SpeechEmotion-2000* (pp. 19-24).
D'Mello, S., Picard, R. W., & Graesser, A. (2007). Toward an Affect-Sensitive AutoTutor. *22*, 53-61.
D'Mello, S., Craig, S., Witherspoon, A., McDaniel, B., & Graesser, A. (2008). Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction, 18*(1), 45-80.
Damasio, A. R. (1996). *Descartes' Error: emotion reason and the human brain*: Penguin Putnam.
de Lera, E., & Garreta-Domingo, M. (2007). Ten Emotion Heuristics: Guidelines for Assessing the User's Affective Dimension Easily and Cost-Effectively *Emotions in HCI workshop, British HCI 2007*.
DeCoster, J. (2004). Meta-Analysis Notes. Retrieved 8 november, 2011, from http://www.stat-help.com/notes.html
Derks, D., Fischer, A. H., & Bos, A. E. R. (2008). The role of emotion in computer-mediated communication: A review. *Computers in Human Behavior, 24*(3), 766-785.
Desmet, P. (2002). *Designing Emotions.* Delft University of Technology, Delft, The Netherlands.
Desmet, P. (2005). Measuring Emotion: Development and Application of an Instrument to Measure Emotional Responses to Products. In M. Blythe, K. Overbeeke, A. Monk & P. Wright (Eds.), *Funology* (Vol. 3, pp. 111-123): Springer Netherlands.
Dreisbach, G., & Goschke, K. (2004). How positive affect modulates cognitive control: Reduced perseveration at the cost of increased distractibility. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(2), 343-353.
Ekman, P. (1993). Facial expression and emotion. *American Psychologist, 48*(4), 384.
Ekman, P., & Friesen, W. (2003). *Unmasking the face: A guide to recognizing emotions from facial expressions.*: Cambridge, MA: Malor Books.
Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin, 128*(2), 203.
Ellsworth, P. C., & Scherer, K. R. (2003). Appraisal processes in emotion. In R. J. Davidson, Goldsmith, H.H. and Scherer, K.R. (Ed.), *Handbook of the affective sciences* (pp. 572–595). New York: Oxford University Press.
Eysenck, M. W. (2004). *Psychology: An International Perspective*. East Sussex: Psychology Press.
Fagerberg, P., Ståhl, A., & Höök, K. (2004). eMoto: emotionally engaging interaction. *Personal and Ubiquitous Computing, 8*(5), 377-381.
Fischer, A. H., & Manstead, A. S. R. (2008). Social Functions of Emotion. In M. Lewis, J. M. Haviland-Jones & L. F. Barrett (Eds.), *Handbook of emotions* (pp. 456-468): Guilford Press.
Frijda, N. H. (1988). The laws of emotion. *American Psychologist, 43*(5), 349.

Frijda, N. H. (2004). Emotions and action. In A. S. R. Manstead & N. H. Frijda (Eds.), *Feelings and Emotions: the amsterdam symposium* (pp. 158–173): Cambridge University Press.
Frijda, N. H., Manstead, A. S. R., & Bem, S. (Eds.). (2000). *Emotions and Beliefs: How Feelings Influence Thoughts*: Cambridge University Press.
Gasper, K., & Clore, G. L. (2002). Attending to the Big Picture: Mood and Global Versus Local Processing of Visual Information. *Psychological Science, 13*(1), 34-40.
George, J. M., Jones, G. R., & Gonzalez, J. A. (1998). The Role of Affect in Cross-Cultural Negotiations. *Journal of International Business Studies, 29*(4), 749-772.
Gilleade, K., Dix, A., & Allanson, J. (2005). Affective videogames and modes of affective gaming: assist me, challenge me, emote me *Proc. DIGRA 2005* (Vol. 2005): Citeseer.
Girard, S., & Johnson, H. (2009). Developing affective educational software products: Sorémo, a new method for capturing emotional states. *Journal of Engineering Design, 20*(5), 493-510.
Graesser, A. C., Chipman, P., Haynes, B. C., & Olney, A. (2005). AutoTutor: an intelligent tutoring system with mixed-initiative dialogue. *Education, IEEE Transactions on, 48*(4), 612-618.
Gratch, J., & Marsella, S. (2001). Tears and fears: modeling emotions and emotional behaviors in synthetic agents *Proceedings of the fifth international conference on Autonomous agents* (pp. 278-285). Montreal, Quebec, Canada: ACM.
Gratch, J., Marsella, S., & Petta, P. (2009). Modeling the cognitive antecedents and consequences of emotion. *Cognitive Systems Research, 10*(1), 1-5.
Ham, W. v. d., & Broekens, J. (2011). *The Effect of Dominance Manipulation on the Perception and Believability of an Emotional Expression.* Paper presented at the Workshop on Standards in Emotion Modelling, Leiden.
Heerink, M., Krose, B., Evers, V., & Wielinga, B. (2006). The influence of a robot's social abilities on acceptance by elderly users *The 15th IEEE International Symposium on Robot and Human Interactive Communication, 2006* (pp. 521-526): IEEE.
Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst., 22*(1), 5-53.
Hone, K. (2006). Empathic agents to reduce user frustration: The effects of varying agent characteristics. *Interacting with Computers, 18*(2), 227-245.
Höök, K. (2008). Affective Loop Experiences – What Are They? *Persuasive Technology* (pp. 1-12).
Hsu, S., Wen, M.-H., Lin, H.-C., Lee, C.-C., & Lee, C.-H. (2007). AIMED- A Personalized TV Recommendation System *Interactive TV: a Shared Experience* (pp. 166-174).
Hudlicka, E. (2003). To feel or not to feel: The role of affect in human-computer interaction. *International Journal of Human-Computer Studies, 59*(1-2), 1-32.
Hudlicka, E. (2008). Affective Computing for Game Design *Proceedings of the 4th Intl. North American Conference on Intelligent Games and Simulation* (pp. 5-12).
Isbister, K., Höök, K., Laaksolahti, J., & Sharp, M. (2007). The sensual evaluation instrument: Developing a trans-cultural self-report measure of affect. *International Journal of Human-Computer Studies, 65*(4), 315-328.
Isbister, K., Hook, K., Sharp, M., & Laaksolahti, J. (2006). The sensual evaluation instrument: developing an affective evaluation tool *Proceedings of the SIGCHI conference on Human Factors in computing systems* (pp. 1163-1172): ACM.
Isen, A. M., Daubman, K. A., & Nowicki, G. P. (1987). Positive affect facilitates creative problem solving. *Journal of Personality and Social Psychology, 52*(6), 1122-1131.
Isomursu, M., Kuutti, K., & Väinämö, S. (2004). Experience clip: method for user participation and evaluation of mobile concepts *Proceedings of the eighth conference*

on Participatory design: Artful integration: interweaving media, materials and practices - Volume 1. Toronto, Ontario, Canada: ACM.
Isomursu, M., Tähti, M., Väinämö, S., & Kuutti, K. (2007). Experimental evaluation of five methods for collecting emotions in field settings with mobile applications. *International Journal of Human-Computer Studies, 65*(4), 404-418.
Janssen, J., van den Broek, E., & Westerink, J. (2012). Tune in to your emotions: a robust personalized affective music player. *User Modeling and User-Adapted Interaction, 22*(3), 255-279.
Janssen, J. H., Bailenson, J. N., IJsselsteijn, W. A., & Westerink, J. H. D. M. (2010). Intimate Heartbeats: Opportunities for Affective Communication Technology. *IEEE Transactions on Affective Computing, 1*, 72-80.
Kellaris, J. J., & Kent, R. J. (1993). An Exploratory Investigation of Responses Elicited by Music Varying in Tempo, Tonality, and Texture. *Journal of Consumer Psychology, 2*(4), 381-401.
Kleef, G. A. v., De Dreu, C. K. W., Pietroni, D., & Manstead, A. S. R. (2006). Power and emotion in negotiation: power moderates the interpersonal effects of anger and happiness on concession making. *European Journal of Social Psychology, 36*(4), 557-581.
Kleinke, C. L. (1986). Gaze and eye contact: A research review. *Psychological bulletin, 100*(1), 78.
Laaksolahti, J., Isbister, K., & Höök, K. (2009). Using the sensual evaluation instrument. *Digital Creativity, 20*(3), 165-175.
Lang, P. J., Bradley, M.M., & Cuthbert, B.N. (2008). International affective picture system (IAPS): Affective ratings of pictures and instruction manual. . *Technical Report A-8. University of Florida, Gainesville, FL.*
Laurans, G., & Desmet, P. M. A. (2006). 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 & J. v. Erp (Eds.), *Design & Emotion 2006*.
LeDoux, J. (1996). *The Emotional Brain*. New York: Simon and Shuster.
Lee, C.-H. J., Chang, C., Chung, H., Dickie, C., & Selker, T. (2007). Emotionally reactive television *Proceedings of the 12th international conference on Intelligent user interfaces* (pp. 329-332). Honolulu, Hawaii, USA: ACM.
Lewis, M., Haviland-Jones, J. M., & Barrett, L. F. (Eds.). (2008). *Handbook of Emotions* (Third ed.): The Guilford Press.
Loewenthal, K. M. (2001). *An introduction to psychological tests and scales*: Psychology Pr.
Markopoulos, P., Read, J., Hoÿsniemi, J., & MacFarlane, S. (2008). Child computer interaction: advances in methodological research. *Cognition, Technology & Work, 10*(2), 79-81.
McQuiggan, S. W., & Lester, J. C. (2007). Modeling and evaluating empathy in embodied companion agents. *International Journal of Human-Computer Studies, 65*(4), 348-360.
Mehrabian, A. (1980). *Basic Dimensions for a General Psychological Theory*: OG&H Publishers.
Mehrabian, A., & Russell, J. A. (1974). A verbal measure of information rate for studies in environmental psychology. *Environment and Behavior, 6*, 233-252.
Ortony, A., Clore, G. L., & Collins, A. (1988). *The Cognitive Structure of Emotions*: Cambridge University Press.
Osgood, C. E. (1966). DIMENSIONALITY OF THE SEMANTIC SPACE FOR COMMUNICATION VIA FACIAL EXPRESSIONS. *Scandinavian Journal of Psychology, 7*(1), 1-30.

Paiva, A. (2000). Affective Interactions: Toward a New Generation of Computer Interfaces? *Affective Interactions* (pp. 1-8).

Paiva, A., Dias, J., Sobral, D., Aylett, R., Sobreperez, P., Woods, S., et al. (2004). Caring for agents and agents that care: Building empathic relations with synthetic agents *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems* (Vol. 1, pp. 194-201). New York, New York: IEEE Computer Society.

Panksepp, J. (1982). Toward a general psychobiological theory of emotions. *Behavioral and Brain Sciences, 5*(03), 407-422.

Panksepp, J. (1998). *Affective Neuroscience: the foundations of human and animal emotions*: Oxford University Press.

Pantic, M., & Rothkrantz, L. J. M. (2004). Facial action recognition for facial expression analysis from static face images. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 34*(3), 1449-1461.

Pantic, M., Sebe, N., Cohn, J. F., & Huang, T. (2005). Affective multimodal human-computer interaction *Proceedings of the 13th annual ACM international conference on Multimedia* (pp. 669-676). Hilton, Singapore: ACM.

Picard, R. W. (1997). *Affective Computing*: MIT Press.

Picard, R. W. (2003). Affective computing: challenges. *International Journal of Human-Computer Studies, 59*(1-2), 55-64.

Picard, R. W., & Klein, J. (2002). Computers that recognise and respond to user emotion: theoretical and practical implications. *Interacting with Computers, 14*(2), 141-169.

Pommeranz, A., Broekens, J., Visser, W., Brinkman, W.-P., Wiggers, P., & Jonker, C. M. (2008). Multi-angle view on preference elicitation for negotiation support systems *Proceedings of the 1st International Working Conference on Human Factors and Computational Models in Negotiation* (pp. 19-26). Delft, The Netherlands: ACM.

Pommeranz, A., Broekens, J., Wiggers, P., Brinkman, W.-P., & Jonker, C. (2012). Designing interfaces for explicit preference elicitation: a user-centered investigation of preference representation and elicitation process. *User Modeling and User-Adapted Interaction, 22*(4-5), 357-397.

Pongrácz, P., Molnár, C., Miklósi, A., & Csányi, V. (2005). Human Listeners Are Able to Classify Dog (Canis familiaris) Barks Recorded in Different Situations. *Journal of Comparative Psychology, 119*(2), 136.

Pouwelse, J. A., Garbacki, P., Wang, J., Bakker, A., Yang, J., Iosup, A., et al. (2008). TRIBLER: a social-based peer-to-peer system: Research Articles. *Concurr. Comput. : Pract. Exper., 20*(2), 127-138.

Read, J. (2008). Validating the Fun Toolkit: an instrument for measuring children's opinions of technology. *Cognition, Technology & Work, 10*(2), 119-128.

Reeves, B., & Nass, C. (1996). *How people treat computers, television, and new media like real people and places*: CSLI Publications and Cambridge university press.

Reisenzein, R. (1994). Pleasure-Arousal Theory and the Intensity of Emotions. *Journal of Personality and Social Psychology, 67*(3), 525-539.

Rolls, E. T. (2000). Precis of The Brain and Emotion. *Behavioral and Brain Sciences, 20*, 177-234.

Russell, J. (2003). Core affect and the psychological construction of emotion. *Psychological Review, 110*(1), 145-172.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161-1178.

Russell, J. A. (1994). Is there universal recognition of emotion from facial expressions? A review of the cross-cultural studies. *Psychological bulletin, 115*(1), 102.

Russell, J. A., & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality, 11*(3), 273-294.

Russell, J. A., Weiss, A., & Mendelsohn, G. A. (1989). Affect Grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology, 57*(3), 493.

Sánchez, J. A., Hernández, N. P., Penagos, J. C., & Ostróvskaya, Y. (2008). Conveying Mood and Emotion in Instant Messaging by Using a Two-Dimensional Model for Affective States *Anais do IHC 2006* (pp. 66-72).

Scherer, K. R., Schorr, A., & Johnstone, T. (Eds.). (2001). *Appraisal Processes in Emotion: Theory, Methods, Research*: Oxford University Press.

Scherer, K. R., & Wallbott, H. G. (1994). Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology, 66*(2), 310-328.

Soleymani, M., & Larson, M. (2010). Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus. In M. L. V. Carvalho, Yilmaz (Ed.), *Proceedings of the SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010)* (pp. 4-8).

Sundström, P. (2005). Exploring the affective loop. *Licenciate Thesis, Stockholm University, Stockholm, Sweden*.

Sykes, J., & Brown, S. (2003). Affective gaming: measuring emotion through the gamepad *CHI'03 extended abstracts on Human factors in computing systems* (pp. 732-733): ACM.

Tunstall, P., & Gipps, C. (1996). Teacher Feedback to Young Children in Formative Assessment: A Typology. *British Educational Research Journal, 22*(4), 389-404.

van Steenbergen, H., Band, G. P. H., & Hommel, B. (2010). In the Mood for Adaptation. *Psychological Science, 21*(11), 1629-1634.

Walther, J. B., & D'Addario, K. P. (2001). The impacts of emoticons on message interpretation in computer-mediated communication. *Social Science Computer Review, 19*(3), 324.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales. *Journal of Personality and Social Psychology, 54*(6), 1063-1070.

Weisfeld, G. E., & Beresford, J. M. (1982). Erectness of posture as an indicator of dominance or success in humans. *Motivation and Emotion, 6*(2), 113-131.

Woolf, B., Arroyo, I., Cooper, D., Burleson, W., & Muldner, K. (2010). Affective Tutors: Automatic Detection of and Response to Student Emotion Advances in Intelligent Tutoring Systems. In R. Nkambou, J. Bourdeau & R. Mizoguchi (Eds.), (Vol. 308, pp. 207-227): Springer Berlin / Heidelberg.

Yannakakis, G. N., & Hallam, J. (2009). Real-Time Game Adaptation for Optimizing Player Satisfaction. *Computational Intelligence and AI in Games, IEEE Transactions on, 1*(2), 121-133.

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 31*(1), 39-58.