

humaine

D9g

Taxonomy of Affective Systems Usability Testing

Workpackage 9 Deliverable



Date: 30th November 2006

IST project contract no.	507422
Project title	HUMAINE Human-Machine Interaction Network on Emotions
Contractual date of delivery	<i>November 30, 2006</i>
Actual date of delivery	<i>December 20, 2006</i>
Deliverable number	D9g
Deliverable title	Taxonomy of Affective Systems Usability Testing
Type	Report
Number of pages	52
WP contributing to the deliverable	WP 9
Task leader	UNIVLEEDS
Author(s)	S.J. Westerman, P.H. Gardner, E.J. Sutherland
EC Project Officer	Philippe Gelin

Address of lead author: Steve Westerman
Institute of Psychological Sciences
University of Leeds
Leeds, LS2 9JT
UK

Table of Contents

Section	Title	Page
1	The place of this report within HUMAINE	5
1.1	The field covered by Workpackage 9	5
1.2	The research objectives	5
1.3	Main elements of the deliverable	6
2	Emotion-related computing systems	8
2.1	Models of affect	8
2.2	The function and consequences of emotions	11
3	Usability testing	14
3.1	The conventional approach to usability assessment	14
3.2	The changing face of usability	16
3.3	Data gathering methods	19
3.3.1	<i>Self-report</i>	20
3.3.2	<i>Non-verbal behaviour</i>	22
3.3.3	<i>Psychophysiological assessment</i>	23
3.3.4	<i>Usability inspection methods</i>	25
4.0	Taxonomic descriptions and models of assessment	27
4.1	The association between affective computing and user experience	27
4.2	Transmission of emotion-related information	28

Section	Title	Page
4.3	Evaluating costs and benefits	32
5	A selective review of affective computing studies	35
5.1	Methods used to assess emotion	35
5.2	Types of task and experimental manipulations	36
5.3	Outcomes	37
5.4	Effects of task difficulty: Old wine in new bottles?	38
5.5	Differentiating emotions	39
6	Conclusions	40
6.1	‘Usability testing’: An appropriate methodology for the evaluation of emotion-related systems?	40
6.2	Methods of assessment	41
6.3	The manipulation of affect	42
6.4	Standardization	42
6.5	Differentiating emotions	43
6.6	Final remarks	44
	References	45

1. The place of this report within HUMAINE

The HUMAINE Network of Excellence (NoE) is an EC supported activity that brings together researchers with interests and expertise relevant to the development of computing systems that can “register, model and/or influence human emotional and emotion-related states and processes” (HUMAINE Technical Annex). This report contributes to Workpackage 9 of this NoE.

1.1 The field covered by Workpackage 9

Workpackage 9 (WP9) is examining processes and methods relating to the assessment of the usability of emotion-related systems. For many years HCI researchers have argued that assessing usability is an important and integral part of the process of computing system development. Awareness has increased in industry and slowly user centred design (UCD) is becoming a more established industrial practice (Vrendenburg et al., 2002; Mao et al., 2005), supported by a variety of tested methods for data gathering and analysis. However, user-centred assessment of emotion-related computing systems presents fresh challenges for HCI researchers. It is not clear how effective or appropriate many of the existing techniques for usability assessment will be in this new arena. WP9 is concerned with furthering understanding in this area, by examining existing methods and, where appropriate, proposing alternatives.

1.2 The research objectives

Much of the planned activity in WP9 (see HUMAINE project deliverable D9b) is based on the utilisation of qualitative methods and formative (see Karat, 1997) usability assessments that can be used as an integral part of a design process. This report is intended to be complementary, expanding the focus of WP9, by giving priority to quantitative assessments of usability and their application to emotion-related computing systems. The objectives of this report are as follows:

- To consider key issues that may influence the selection and application of quantitative usability methods in the context of emotion-related computing systems.

- To develop simple taxonomies/frameworks/models that can be used as a basis for consideration of testing emotion-relating systems in this report and beyond.
- To evaluate existing research evidence in this context.

When considering research evidence for inclusion in this report, emphasis is placed on ‘body measures of affect’ (Picard and Daily, 2005), including physiological and behavioural assessments. Psychometric assessments are considered in Workpackage deliverable D9f.

The report is not intended as a primer. However, it is recognised that some of the audience will not be familiar with all topic areas addressed. For this reason some background material is included where appropriate.

1.3 Main elements of the deliverable

By way of ‘setting the scene’ we begin this report (Section 2) with a brief overview of models of emotion and affective states. The differing perspectives that are presented in the research literature have implications for the design of emotion-related computing systems and assessments of their usability. We also consider the adaptive significance of emotions, as this may provide useful indications when discussing the benefits that can be expected from emotion-related systems. Following this (Section 3) issues relating to usability are examined. Key features of usability assessments as applied to ‘conventional’ systems are described and the extent to which existing methodologies and the underlying theoretical principles are relevant to emotion-related systems is considered. In Section 4 we describe some simple frameworks that can be applied to emotion-related computing systems. These are used as a basis for developing a better understanding of system and user requirements and limitations. *Emotion-related computing systems* is a very broad domain and there are important differences in included systems. For example, a useful distinction can be made between those systems that provide a dynamic response to the users’ emotional state (as inferred from, e.g., biometrics or key logging data), and those that consider User Experience (UX) but do not adapt (Hassenzahl and Tractinsky, 2006). The need for an appraisal of costs and benefits associated with system components is also discussed. In Section 5, we review some relevant primary research. This is restricted to studies that include physiological measurement, that focus on computer-based performance (rather than abstract laboratory tasks), and that have manipulated or monitored users’ affective experience in such a way that contrasts can be made between experimental conditions. Finally, in Section 6, some conclusions are drawn

regarding the assessment of emotion-related computer systems, and items for the research agenda are considered.

The following people have contributed to the work reported in this deliverable:

S.J. Westerman, P.H. Gardner, & E.J. Sutherland

The institutions that have contributed are:

University of Leeds

2. Emotion-related computing systems

Emotion is a ubiquitous part of human existence. For example it influences cognitive processing (Isen & Daubman, 1984; Isen, Daubman and Nowicki, 1987), is concerned with preparation for action (Lang, Bradley & Cuthbert, 1997; Frijda, 1986), and is involved in communication (Banse and Scherer, 1996). On this basis, it seems intuitively plausible that the efficiency of computer systems may be greater, or the experience of the user may be more positive, if system design takes account of user emotion, and systems are able to recognise and respond dynamically. This is the central premise behind research activity in the area of *affective computing* (see Picard, 1997) with, for example, many studies trying to identify reliable associations between psychophysiological measurements and the process of interaction (e.g., Ward, 2004). More generally, the affective experience of the user has become a topical issue for the HCI research community (see e.g., Cockton, 2004; Ward & Marsden, 2004; Hassenzahl & Tractinsky, 2006) and has prompted some to propose revisions to existing approaches to usability assessment (e.g., Dillon, 2001). The evidence to support this position is discussed in subsequent sections of this report.

For readers who require more detailed information, Picard (1997) is a seminal text on affective computing, and reviews can be found in Hudlicka (2003), Picard & Klein (2002), Picard (2003), and McNeese (2003).

2.1 Models of affect

A clear understanding of the nature of affect and emotion would seem to be a prerequisite for successful research in this area. However, currently the research community is divided on many details. In both the Psychology and Computer Science literatures, the terms ‘affect’ and ‘emotion’ are used in different ways by different authors. One view is that ‘affect’ is a broad (and consequently less precise) term that encompasses more specific (and therefore more detailed) terms such as ‘emotion’ and ‘mood’. For example, Isen (2004, p. 264) considers affect to encompass unconscious processes and “cognitive, neurological, physiological, motivational, and behavioural components, as well as the feeling component” (see also Dillon, 2001). In contrast, Power (2006, p. 694) regards affect as “the conscious experience of emotion”. Russell (2003) proposes another relationship, with ‘core affect’ being a fundamental property that contributes to emotions and moods that possess other facets. Still others, including many writing on the topic of affective computing (see e.g., Dillon, 2001;

Picard, 1997), seem to use the terms ‘affect’ and ‘emotion’ interchangeably (Power, 2006). In this report, affect is recognised as a broader, more generic, term. However, because it has been argued that ‘affective computing’ systems form only a subset of those systems that deal with affective responses of the user (Hassenzahl & Tractinsky, 2006), and for consistency with other HUMAINE reports, in this document the term ‘emotion-related’ will be used to refer to the domain. When considering *affective computing systems* further qualification is required, and this is addressed below.

The structure of emotion has long been a subject of contention. Broadly, two contrasting views can be identified. There are those who describe emotion with reference to a small number of dimensions, such that an individual’s current state constitutes a point in multi-dimensional affective space (Russell, 1980, 2003; Scherer, 2005; Watson & Tellegen, 1985). Others believe that a number of basic emotions exist such that categorisation of fundamental types of emotional is appropriate (e.g., Ekman, 1999; Izard, 1977). From both perspectives, important details remain to be resolved. Among subscribers to the ‘dimensional’ view there is some consensus on a two-dimensional affective space (Russell, 1980; Watson & Tellegen, 1985) (see Figure 1). Nevertheless, the cardinal dimensions of affective space are the subject of continued debate (Carver, 2004). Some argue that pleasure and arousal constitute major orthogonal dimensions (e.g., Russell, 2003), others advocate positive affect (PA) and negative affect (NA), as dimensions oriented at 45 degrees to pleasure and arousal (e.g., Watson & Tellegen, 1985). PA and NA have been linked to approach and avoidance behaviours, respectively. However, Watson, Wiese, Vaidya, & Tellegen (1999) suggest that neither of these pairs of dimensions are orthogonal, finding pleasure and activation to be positively correlated and PA and NA to be negatively correlated. They also question the bipolarity of all but the pleasantness dimension. Carver (2004) takes a somewhat different dimensional perspective in presenting evidence to support the contention that the Behavioural Approach System (BAS) and Behavioural Inhibition System (BIS) can both lead to positive and negative affective states depending on the success, or otherwise, of the intended activity (approach or withdrawal). From this, it is possible to explain associations between negative affective states, e.g., frustration, and approach behaviours. This seems particularly relevant to emotion-related computing systems, and we return to this point later.

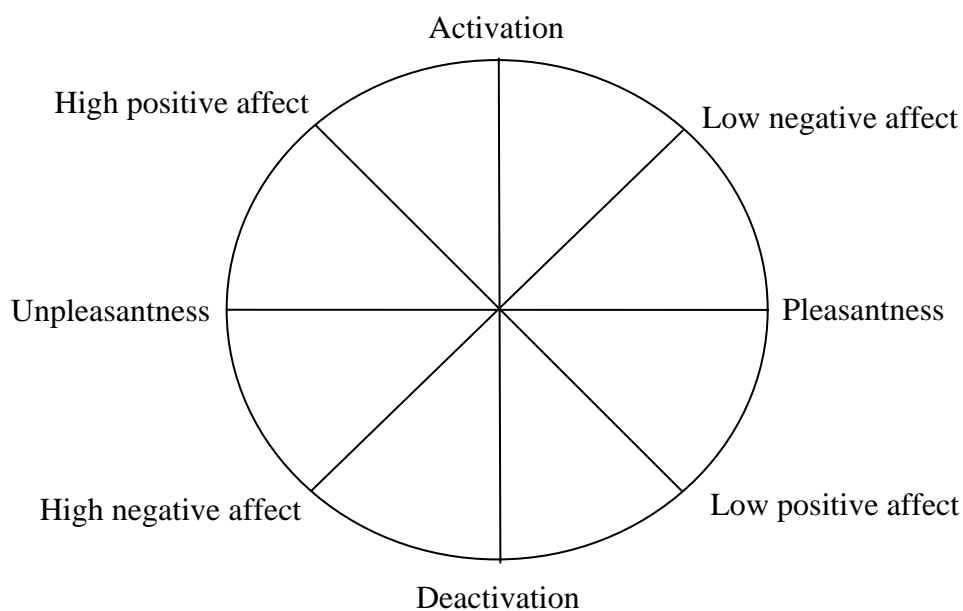


Figure 1. Circumplex of affect (see Watson & Tellegen, 1985; Russel, 2003).

Of those who believe an emotional typology is appropriate, there is general agreement as to what some of these ‘basic’ emotions might be. For example, sadness, happiness, disgust, anxiety, and anger appear in most typological models (Power, 2006). However, there is also disagreement on fundamental issues such as the number of ‘basic’ emotions, with some of the most influential models describing six (Ekman, 1982), eight (Plutchik, 1962, 1980), and ten emotions (Izard, 1971, 1977). Even then, others question whether each of these basic emotions is sufficiently detailed. For example, Russell (2003) makes the point that there may conceivably be a number of different types of fear. In a similar vein, some researchers on the topic of emotion-related computing and affective product design consider secondary emotions to be an appropriate level of description (e.g., Desmet, 2003).

Brain imaging techniques may provide a means of resolving some of these differences. On the basis of results to date, no conclusions can be drawn (Barrett & Wager, 2006). However, it is also possible that strong contrasts between the dimensional and typological positions are illusory. For example, Russell (2003) suggests that the appearance of an emotional typology may result from the cognitive appraisal of what he refers to as ‘core affect’ – affective state as described by the two-dimensional circumplex. Cognitive appraisal is an important element of many current models of emotion (for example Ekman, 1977; Lazarus, 1982; Scherer, 2005), and for some researchers in the area of emotion-related

computing (e.g., Bessiere et al., 2006) what can be referred to as pre-emotional states are a major focus of interest.

Obviously, when considering emotion-related computing systems, it is important to be able to place research in the context of these broader models of affect. We return to this topic in later sections. In some instances researchers have translated their results to allow consideration of implications for both dimensional and typological models (e.g., Picard et al., 2001; Mandryk et al., 2006).

2.2 The function and consequences of emotions

Although affect is such a prominent feature of human experience, this should not be taken to signify that system design features that take account of emotion will enhance all aspects of our interaction with computers, or, indeed, any. In an interesting and thought-provoking paper that addresses the rationale for including consideration of emotions when designing computer systems, Bellman (2002, p. 150) poses the rhetorical question: “Why shouldn’t we leave the topic [emotion] alone as a human – or at least animal – phenomenon, to be discussed by psychologists and physiologists but certainly not engineers and technologists?” With this question in mind, consideration of the adaptive significance of emotions and the role they play in human functioning provides useful information as to the benefits that can be anticipated and the types of computing contexts in which these might accrue. This exercise forms part of an appraisal of costs and benefits, that will inform the process of usability assessment. Implementing marginally useful features would not be sensible if the costs associated with doing so were disproportionately high (see Section 4.3). There are three, somewhat related, features of emotion that we consider in this context (cf., Hudlicka, 2003). The first is the involvement of emotion in the regulation of appetitive and avoidance behaviours (Lang, Bradley and Cuthbert, 1997). The second is the influence of emotion on cognitive performance and possibly cognitive capacity (Isen, Daubman and Nowicki, 1987). The third is the role of emotion in social interaction and bonding (Lopes, Salovey, Cote, Beers & Petty, 2005).

There are obvious adaptive reasons for avoiding threatening, harmful situations. Organisms will lead longer and healthier lives if they do so. In this context the protective role of emotions such as fear can readily be seen. However, adventuresome, exploratory behaviour also has adaptive significance. Humans would make little, if any, progress if their only

concern was avoiding harm. Emotions play a key role in both ‘poles’ of this motivational system. An adaptive function of emotions may be the facilitation of situation appraisal making the processing of relevant information more efficient (Bellman, 2002; Reeves & Nass, 2002; Curtis, Aunger and Robbie, 2004). Although, in the context of human-computer interaction situations of extreme fear or extreme desire happen rarely, if at all, appetitive and avoidance behaviours, and related, albeit less intense emotions, have relevance. If interaction with a computer system produces an aversive emotional reaction, such as annoyance or frustration, continued interaction becomes less likely. Following from Carver (2004: see above) frustration (a negative affective state) can be the result of ineffective approach behaviour, i.e., the user not being able to achieve a goal. Interaction is also likely to be hampered by restricted cognitive processing resulting from negative affect (Isen, 1985) that may be detrimental to efforts to find alternative, and more productive, approaches (Norman, 2004). Conversely, if an emotion-related computer system is able to anticipate or detect such negative affective states this should provide an important step towards the avoidance of these unwanted outcomes. Further differentiation of emotions (beyond simple appetitive/aversive reactions) may be useful in providing the individual with information about the nature of the evoking stimulus (Scherer, 2005).

Emotions are an integral component of human-human interaction (Parkinson, 2005). The expression and interpretation of emotion in this circumstance can take several forms (e.g., tone of voice, posture). Effective use of emotional information facilitates social adjustment (e.g., Engelberg and Sjoberg, 2004) and contributes to the processes of social cohesion (Lawler, 2001). The concept of ‘emotional intelligence’ has been proposed to describe human abilities relating to the conveyance and understanding of emotions (see e.g., Goleman, 1995). Individuals who are adept at interpreting emotion and conveying appropriate emotions are held to be more emotionally intelligent. The assumption is that they are more successful in adapting to their environment. Emotional communication can form the basis for inferences about the characteristics of others and result in appetitive/avoidance behaviour in the individual. For example, we may consider someone who apparently disregards our emotions as unfriendly or ‘cold’ and choose to avoid them. Importantly, there may be times when it is appropriate or desirable to withhold displays of emotion or to display an emotion that we are not experiencing. We do not always ‘wear our hearts on our sleeves’.

Examining this communicative aspect of emotion from the perspective of the usability of human-computer systems, there are two key points to be considered. First, do advantages

accrue if human-computer communication mirrors human-human communication? Reeves & Nass (2002) present a good deal of evidence to support their premise that humans view interacting with computers very similarly to interacting with other humans. This might suggest that emphasising parallels will be valuable. However, there are other issues to consider. In human-human communication there is no alternative than to recognise the role of emotion. When interacting with a computer there is a choice. Communication of emotion relates to social bonding (Lopes, Salovey, Cote, Beers & Petty, 2005) and this may not be an important consideration in human-computer interaction. Nevertheless, users might enjoy an interaction process in which computer responses are related to affect, but will this be a more effective method of interaction and, if not, should enjoyment or effectiveness be given priority? These are important questions to address as part of the appraisal of emotion-related computing systems.

The second point concerns the potential for the user to communicate an affective state to the computer other than the one they are actually experiencing. The user may not want to divulge an experienced emotion, may want to emphasise an experienced emotion, or may want to communicate a completely different emotion in the hope of eliciting a particular response from the computer, or others engaged with the system. It is interesting to note that it is relatively complex to implement computer-based assessment of the means of detecting emotions that humans find most effective (e.g., facial muscle movements), and that humans are relatively poor at detecting differences relating to measures that are fairly straightforward to implement on computer (e.g., galvanic skin response: GSR).

3. Usability testing

In this section of the report we begin by describing ‘conventional’ approaches to the assessment of usability. This provides a basis for subsequent consideration of a broader conceptualisation that better accommodates user motivation. Following this, key methods of gathering data are described and evaluated with regard to a set of proposed criteria (based on O’Donnell & Eggemeier, 1986) and suitability for assessing emotion-related computing systems is discussed.

A comprehensive treatment of all aspects of usability is not attempted here. A useful review of usability testing can be found in Hornbaek (2006) and more general descriptions and discussions of the topic in Nielsen (1993) and Wixon & Wilson (1997).

3.1 The conventional approach to usability assessment

In some form, ‘usability’ has been recognised as an important topic for human-machine interaction research for nearly 50 years, with a significant body of work referenced by descriptors such as ‘ease of use’ and ‘ergonomics’ (see Hornbaek, 2006). The uptake of the concept and associated methods by industry is progressing, but there is still some distance to travel. Issues such as the availability of trained practitioners (Gulliksen et al., 2004) and the potential for cost/benefit justifications (Vredenburg et al., 2002) are key factors to be addressed.

For the practitioner, evaluating usability involves the development of an understanding of the characteristics of the user, the task, and the context. Usability analyses and software requirements are built on this basis. Various methods of task analysis are available (see e.g., Diaper, 1990). However, identifying a representative set of tasks for usability testing is a persistent difficulty (Ivory & Hearst, 2001). This is particularly relevant to the study of emotion-related computing systems, for which appropriate manipulations of emotion must be considered.

The term *usability* implies a single construct, but is generally held to be multi-faceted, and possibly multi-dimensional (see below). Perhaps the most influential definition is provided in ISO 9241-11 (ISO, 1998, p. 2):

[The] “Extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.”

These component elements are further defined as follows:

- Effectiveness: *How accurately and completely users are able to perform their specified goals.*
- Efficiency: *The amount of effort that is required to achieve the level of effectiveness in performance of the goals.*
- Satisfaction: *A lack of discomfort, and a positive attitude towards the system, while performing the goals.*

Although few will argue with the importance of these components, additions have been proposed, including ‘learnability’ and ‘memorability’ (Nielsen, 1993) and accessibility, universality, and risk (Bevan, 2006).

The degree of association between components has been the subject of some debate. Frøkjær, Mertzum, & Hornbæk (2000) present evidence to support the position that efficiency, effectiveness, and satisfaction are relatively independent. More recently, Sauro & Kindlund (2005) came to the opposite conclusion, although on the basis of relatively modest correlations (in the range .15 to .59). Sauro & Kindlund (2005) also present evidence to suggest that each of these components contributes in relatively equal measure to a single usability factor, while Frøkjær *et al.*, (2000) found the relative importance of each dimension varied depending on the task and the context, e.g., whether tasks are routine or non-routine. This complex combination of results could, at least in part, be explained by non-linear associations between component elements. Jordan (2000) takes the view that, in product design, most consumers will expect products to be easy to use. In this respect, he makes little distinction between physical products (such as training shoes and can-openers) and computer interfaces. The logical extension of this argument is that computer users may not be surprised that an interface is easy to use, but they will display an adverse reaction if it is not easy to use (see Zhang & von Dran, 2000). Consistent with this, Cheung and Lee (2005) found that negative usability assessments had a stronger effect on ratings of satisfaction with websites than positive usability ratings. Correlations between components may vary depending on their position on a positive/negative continuum.

A human bias for negative affect (Frijda, 1988; Reeves & Nass, 2002) may also be important in this regard. Studies have demonstrated that people tend to prioritise negatively valenced events (see Reeves & Nass, 2002). They pay more attention to them and are better able to remember material that follows negative events. This being the case, it would be sensible for primary concern, when designing human-computer interfaces, to be the avoidance of negative events (e.g., avoiding frustration that arises from poor design) and, only when this has been achieved for attention to be switched to identifying potential for positive outcomes. This can be seen to be consistent with conventional treatments of usability that are geared to the avoidance of negative affective responses, including dissatisfaction. Results from a gaming context suggest that there must be exceptions to this position. Johnson & Wiles (2003) conclude that, in a gaming context, positive affect can be generated by interface design elements that serve to reduce usability as it is traditionally conceived (e.g., withholding information from the user and allowing the user to make errors). However, this might be attributable to an experience of positive affect that arises from feelings of achievement under difficult circumstances.

3.2 The changing face of usability

Relatively recently, it has been argued (see e.g., Dillon, 2001; Johnson & Wiles, 2003) that this conventional view of usability is too restricted, with its emphasis on performance-related assessments and the avoidance of dissatisfaction (ISO-9241). It may have arisen due to a primary focus on studying the usability of computers in work contexts (Hassenzahl et al., 2000; see also Mandryk, Atkins, & Inkpen, 2006). The result is that usability assessments say little about user motivation. With more *discretionary use* software becoming available (e.g., web pages) and higher general standards of effectiveness and efficiency, broader consideration of affect is required (e.g., Dillon, 2001; Hassenzahl et al., 2000; Jordan, 2000). For some software (e.g., games) achieving positive affective states is a primary user goal (Johnson & Wiles, 2003). Whether such changes should be regarded as a revision of the existing concept of usability or as a progression to something qualitatively different is debatable. In our view the latter is the more tenable position. Some now favour the term *user experience* (see Hassenzahl & Tractinsky, 2006) and refer to *designing user experiences* (Hassenzahl et al., 2000, p. 202).

From this perspective Dillon (2001; see also Hornbaek, 2006) proposes that usability assessment should now consider: i) outcomes; ii) process; and iii) affect. Outcomes refer to

“what the user attains from the interaction” (p. 61), and relates to ‘traditional’ measures of completion, but can also include products of task performance (e.g., items retrieved from a database). ‘Process’ refers to the process of interaction. This can incorporate considerations of efficiency of performance, but relates more broadly to factors such as interaction sequences, learning, changes in user motivation, and access to help facilities. Finally, ‘affect’, refers to the wider range of user emotion-related responses that were mentioned above. It includes consideration of factors such as response to the aesthetics of the system interface and user frustration.

In the same way that associations between the component elements of the ‘traditional’ model of usability have been examined, paths of influence between the elements in this new model can be hypothesised (see Figure 2).

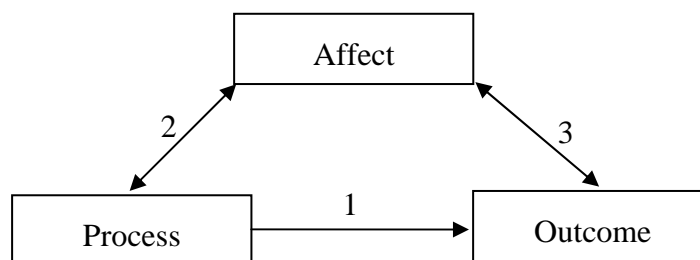


Figure 2. Hypothesised ‘paths of influence’ between Process, Outcome, and User Affect

The numbered arrows in Figure 1 denote associations between elements, as follows:

- 1 The user engages in a process that is designed to achieve a goal, i.e., produce an outcome.
- 2 a Engaging in this process can produce a change in the user’s affect. For example, the

user becomes frustrated because of lack of feedback from the interface. The level of cognitive demand imposed by the task can also impact the user’s affective state (see Section 3.3.3).

- b The user’s affective state can influence the process of interaction. For example, a user who is feeling interested in the material they have retrieved using a keyword search of a database may decide to browse the database and explore further.
- 3 a The user’s affective state can be a desired outcome. For example, feelings of enjoyment and/or excitement might be an appropriate outcome from playing a computer game.
- b The user’s affective state can be influenced by the outcome. For example, the user might be pleased with a drawing they have produced on the computer.

Deciding on appropriate measurements for these constructs is an important concern. As described above, measures of outcome will relate to effectiveness of performance. They will include assessment of task completion and, in the broader sense, assessments of goal achievement. Measures of ‘process’ will relate to performance efficiency and will include measures such as speed and cognitive demand. However, they will also relate to performance strategy, e.g., sequences of actions. Finally, a new conceptualisation of user affect may be required. A possible two dimensional representation that incorporates many important affective constructs as they apply to HCI can be found in Scherer (2005) (see Figure 3).

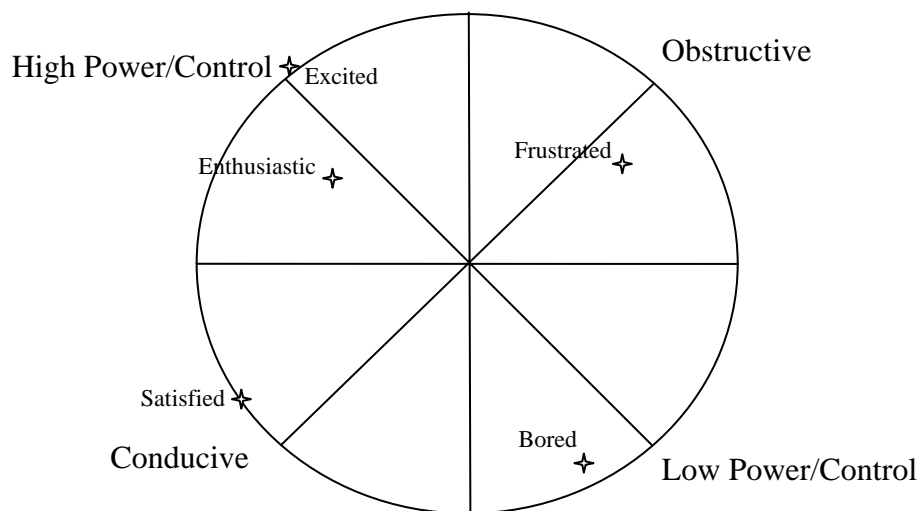


Figure 3. Scherer's (2005) alternative dimensional structure for semantic space for emotions, including some key HCI terms.

3.3 Data gathering methods

Assessment of the constructs set out in ISO 9241 only requires a small number of data gathering methods (Bevan, 2006). However, a wide range of methods is available to researchers and practitioners providing the means to support detailed examination of the implications of specific software designs and guide all stages of the design process. As conceptions of usability assessment change (see above), these methods need to be adapted and supplemented. For example, psychophysiological measurement techniques are now being used in ways that are novel for HCI research (see below), and several psychometric studies have begun to identify affective components of user experience (these are reviewed in deliverable D9f). All methods have their strengths and weaknesses and generally a multi-method approach to the assessment of usability (user experience) will produce the most reliable results (Henderson et al., 1995; Ivory & Hearst, 2001)

In this section of the report we provide a brief overview of key data gathering methods for the assessment of 'traditional' usability and consider potential strengths and weaknesses in the context of emotion-related systems. To support a critical evaluation, we draw on the work of O'Donnell & Eggemeier (1986) who, when reviewing methods for detecting operator workload, identified five key criteria against which they could be assessed: i) sensitivity; ii) diagnosticity; iii) intrusiveness; iv) implementation requirements; and v) operator acceptance. We have adapted these, to make them relevant to the consideration of emotion-related computing systems, and added one further criterion that is particularly appropriate, *temporal resolution* (see Table 1). If an emotion-related system is to adapt to the affective state of the user, it is important that it is able to do so at a pace that is synchronous with the pace of interaction. Equally, it is important that usability assessments are able to evaluate this temporal information.

Criterion	Explanation
Sensitivity	Capability of the technique to discriminate significant variations in

	emotion resulting from task or interface manipulations
Diagnosticity	Capability of a technique to discriminate different dimensions or types of emotion on the basis of task or interface manipulations
Temporal resolution	Capability of a technique to produce an accurate temporal record of changes in the user's emotional state.
Task intrusion	The tendency for a technique to conflict with task performance
User intrusion	Extent to which users feel comfortable being assessed using the particular assessment technique.
Implementation requirements	Factors relating to the ease of implementing a technique, e.g., equipment requirements.

Table 1. Categories of assessment for usability techniques as applied to emotion-related systems (adapted from O'Donnell & Eggemeir, 1986).

3.3.1 Self-report

There are several methods of data gathering based on users' self-report. Some provide data that are amenable to quantitative analysis, e.g., questionnaires that use a standardized rating scale response format. Others provide data of a more qualitative nature (e.g., focus groups). These qualitative methods can be particularly valuable in the early stages of the design process as part of 'scoping' the problem. Questionnaires are available that have been designed specifically for the purpose of making global (and sometimes detailed) assessments of system/interface usability (see e.g., Gedgia, Hamborg, & Duntsch, 1999; Kirakowski & Corbett, 1993). The structure of some of these questionnaires reflects the approach to usability described in ISO 9241 (see above). Some are commercial products, others are freely available. Most have been developed using psychometric techniques and assessed for reliability and validity. Many psychometric studies have examined dimensions of affective response during computer-based performance. However, consistency of approach has been lacking, making interpretation of results difficult (Hornbaek, 2006). A detailed evaluation of key questionnaires is provided in HUMAINE deliverable D9f.

Interviewing users is a self-report technique frequently applied during system development. Semi-structured interview schedules are often favoured as they facilitate

comparison of comments made by participants and the aggregation of data. Focus groups represent a related technique that allows participants to share ideas so that discussion on new topics can be stimulated. This is a time-efficient method enabling views from many participants to be obtained quickly. However, focus groups can be more susceptible to certain types of response bias (e.g., social desirability). Moreover, in the context of emotion-related systems, individual differences in response may be sufficient to make any tendency toward group responses a more detrimental feature. Finally, verbal protocol is a technique that requires the user to provide a ‘running commentary’ accounting for their actions when using a piece of software under development. This can provide the researcher with insight into ‘hidden’ cognitive (and potentially emotional) processes. It can provide valuable results identifying usability issues (Henderson et al., 1995). However, analysis of protocols can be time consuming and generating protocols can be a difficult task for users, running the risks of: i) changed task performance; and, ii) users rationalising their behaviour.

Self-report methodology has advantages with respect to the ease with which data can be gathered. It is cheap to apply and has high face validity. However, self-reports are susceptible to bias of various types. Most importantly, not all relevant aspects of cognition and emotion are open to conscious introspection (see e.g., Nisbett & Wilson, 1977). When considering the potential for application of this method in the context of emotion-related computing systems, the diagnosticity of self report would be expected to be good, and different emotional dimensions and types are detectable (e.g., Power, 2006). However, temporal resolution is generally poor, in that most methods of self-report are not readily time-shared with task performance, and therefore tend to be used to assess periods of interaction. Alternative methods have been developed that allow data gathering to coincide with task performance (e.g., Conati, 2004; Cowie et al., 2000). This improves temporal resolution but data gathering becomes very intrusive of task performance.

3.3.2 Non-verbal behaviour

Monitoring users’ task-related behaviour and performance can take many forms. It can be conducted in controlled, laboratory environments or in the field. It can involve taking notes of observed behaviours, recording behaviours through video analysis, asking users to maintain diaries, or computer logging of user input (e.g., keystrokes, mouse movements) and system state (see e.g., Okada & Asahi, 1999). Non-verbal methods for detecting the emotional state of the user are somewhat different in nature from those used for ‘traditional’ usability

assessment, and include assessments of user posture, voice, and gestures. For example, Kapoor et al. (2004) recorded posture, using pressure sensors on a chair, as part of the input into a multi-modal assessment of user affective state.

For assessment of the ISO 9241 components of usability, measurements of speed (efficiency), accuracy (effectiveness), and task completion (effectiveness) would be usual. Self-report assessment of workload is sometimes used as a supplementary index of efficiency. However, task specific variations are possible. For example, in the context of information retrieval, measures of precision and recall can be valuable (see Baeza-Yates & Ribero-Neto, 1999). These provide an index of the effectiveness with which items of information contained in a database are identified as relevant. When assessing user performance, background (software) logging of user activity provides a high degree of accuracy and the option of more detailed analysis. For example frequency of transitions between specific software states can be examined (e.g., Smith, 1996).

Logging user key strokes and other input activity is potentially a very unobtrusive method of data gathering. However, large quantities of data can be produced and these may need disambiguating by cross-referencing to other data. For example, Joachims et al. (2005) paired navigation patterns with eye tracking data (see below) in an analysis of information retrieval; and Lin & Imamiya (2006) found the combination of eye tracking and mouse movement data useful for user performance evaluation. Other methods of reducing data processing load include the use of parsing routines on the basis of specific system states (see e.g., Kapoor et al., 2004). As far as we are aware there is little research evidence directly addressing the use of patterns of interaction to predict user affective state. However, in support of the possibility, Scheirer et al. (2002) identified four different patterns of mouse clicking in response to frustrating events; and a related example is provided by Wensveen, Overbeeke, & Djajadiningrat (2002) who developed an interface for an alarm clock that enabled interaction to be a basis for judging the affective stage of users.

Eye tracking is becoming much more widely used as a means of usability assessment. As technology progresses eye trackers are much less intrusive. Modern equipment uses infrared cameras to track the movement of the pupils and does not require that the user's head remain motionless. Gaze duration can provide an index of the visual attention of the user (e.g., Prendinger et al., 2005) and tends to be determined by the amount or novelty of information contained at a particular spatial location within the visual scene (see Rayner, 1998). In the

context of HCI, Lin & Imamiya (2006) found number of fixations and scanpath length to be sensitive to manipulations of computer game difficulty. There is some evidence that the emotional content of the visual scene is also a determinant of gaze, although the nature of this effect may be influenced by the age (Isaacowitz et al., 2006) and personality (Isaacowitz, 2006) of the perceiver, with older adults and optimists tending to direct their visual attention away from unhappy images (faces).

3.3.3 Psychophysiological assessment

Human Factors practitioners have been using physiological measures since the late 1970s. A range of measures have been examined as possible indices of operator workload in the context of complex systems (e.g., air traffic control). These include measures of ANS activity, such as galvanic skin response (GSR), measures relating to heart rate (HR), pupil diameter, and measures of respiration; and also measures of cortical activity, including electroencephalograms (EEGs) and evoked potentials (ERPs). The emphasis has been on the detection of stress states with a view to being able to design systems that avoid operator overload. Promising measures include heart rate variability (HRV), the P300 component of evoked potentials, and pupil diameter (see O'Donnell & Eggemeier, 1986; Kramer, 1991, for reviews). Similar techniques have also been explored in the context of 'traditional' human-computer interaction, although the literature is more limited. For example, Lin & Imamiya (2006) found HRV to be sensitive to a manipulation of computer game difficulty. Wilson & Sasse (2000; Wilson, 2001) report changes in HR, GSR, and blood volume pulse (BVP) resulting from degradations of media quality (video frame rates and audio quality), and Iqbal et al. (2005) found pupil diameter to be sensitive to load manipulations for route planning and document editing tasks and also to sequence of task execution as defined by a hierarchical breakdown of task components.

Interest in the association between psychophysiological measures and 'traditional' methods of assessing usability continues (e.g., Lin, Hu, Omata, & Imamiya, 2005). However, changing views on the importance of the user's affective state (as described above), and a substantial literature on the use of physiological measures to assess emotion (see e.g., Cacioppo et al., 2000), has led to a wider application of these measures in relation to HCI. These methods are now being used in the context of affective computing to assess affective state of computer users. However, it is possible that there is a good deal of overlap in the psychophysiological findings from these two areas (workload and affective computing). Both

are concerned with detecting a negative, stressful state in the user. Many of the manipulations used in affective computing studies to induce frustration could be considered to increase the workload of the user, and some explicitly change cognitive demand (see Section 5 for more detail). There are, however, some interesting differences between the two areas of application. For example, HCI research has tended to investigate HR as a predictor of the state of the user (see section 5), whereas HRV has proved the more successful index of operator workload (Kramer, 1991). Facial EMG has been used in many studies to assess emotion (see e.g., Cacioppo et al., 1990) and has been applied with some success in the context of affective computing (see Section 5). So far as we are aware, these measures were not used in Human Factors studies of workload. Conversely, EEG and ERPs have been used extensively to examine operator workload but have not been used in the context of affective computing.

A substantial advantage of psychophysiological measures is that they provide continuous monitoring of user state and, usually, are not disruptive of task performance (although recording of baseline periods can be an issue). In these respects they provide a strong contrast with self-report methods. When coupled with a time-stamped record of users' activities, this can provide a powerful tool for assessing many aspects of user experience as they relate to specific features of the software. However, psychophysiological measures tend to be sensitive to uncontrolled environmental variations, such as changes in heating or lighting.

A further difficulty with several physiological methods of assessment (e.g., GSR) is that they provide information on arousal but not valence (Ward, 2004). Given the earlier discussion on user motivation, this may not be sufficient, and means that they need to be supplemented with other measurement techniques. Most physiological measures are also relatively intrusive, insofar as they require electrodes to be placed on the user. Although less intrusive methods of gathering physiological data are being developed. These include, for example, the use of sensors embedded in an office chair to detect heart rate (Anttonen & Surakka (2005); sensors in glasses to detect facial muscle activity (Scheierer, Fernandez, & Picard, 1999); sensors in a computer mouse to collect measures of skin temperature, GSR, and heart rate (Crosby et al., 2001), and the use of thermal imaging to detect difference in blood flow in the face that relate to muscle activity (Puri et al., 2005). Ethical issues must also be considered in this context as users have very little control over the responses that are being recorded by some of these measures (cf. Reynolds & Picard, 2004).

3.3.4 Usability Inspection Methods

The cost of running usability studies based completely on test data gathered from users can be prohibitive. It has been argued that a cheaper and more pragmatic solution is to get judges to evaluate the software from a usability perspective (Virzi, 1997). Heuristic Review (HR) is one such technique (Nielsen & Molich, 1990). For this purpose, judges do not need to be usability experts, although the performance of experts tends to be better (Nielsen, 1992). There are advantages for providing a standard set of heuristics against which the software will be judged (Law & Hvannberg, 2004). Sets of guidelines (e.g., Smith & Mosier, 1986) are often too extensive to be easily dealt with in this context. For example, the study by Nielsen & Molich (1992) used the following nine heuristics:

- Simple and natural dialogue
- Speak the user's language
- Minimise user memory load
- Be consistent
- Provide feedback
- Provide clearly marked exits
- Provide shortcuts
- Good error messages
- Prevent errors

Inter-judge reliability can be poor, so there are advantages for using a number of independent judges. Between 3 and 5 has been suggested as optimal (Nielsen & Molich, 1990; Nielsen, 1993). The technique also has the advantage that it can be applied on software at all stages of development, even during the conceptual stage. Concerns have been expressed about the overall effectiveness of the technique (Law & Hvannberg, 2004). One possible approach to improve performance involves developing sets of heuristics that are better 'tuned' to the software under evaluation, and recently researchers have begun to do just this. For example, Sutcliffe & Gault (2004) have developed a set of 12 heuristic to guide the evaluation of virtual reality applications, and Mankoff et al. (2003) developed a set of eight heuristics for use with ambient displays. When considering emotion-related computing, in many areas understanding of the 'mechanics' of the domain is at an early stage. However, a useful

research goal would be to develop a set of heuristics that could support the evaluation of these systems.

4.0 Taxonomic descriptions and models of assessment

In this section of the report we present two models that form the basis for an examination of issues relating to the assessment of effectiveness of the design of emotion-related computer systems. We also discuss the potential for evaluating the costs and benefits associated with affective computing systems.

4.1 The association between affective computing and user experience

Hassenzahl and Tractinsky (2006) draw an interesting distinction between *affective computing* and software design that relates to *user experience* (UX). In their view, the former implies an interface that will adapt, to some degree, to the state of the user and possibly the context of use. The latter has no such requirement, but is designed with an explicit concern for the affective experience of the user. Of course, this also implies that the former will assess the affective state of the user. Highly adaptable systems, by definition, do not make assumptions about the emotional state of the user, but try to ascertain that state by gathering evidence. Interfaces with no adaptability must operate on the basis that all users are in the same emotional state when they perform a particular action. From the perspective of assessing emotion-related computing systems the value of adaptivity depends on whether beneficial interface changes can be identified. Between these two extremes are systems that can adapt to more 'generic' context. For example, Bickmore & Mauer (2006) found that participants developed stronger social bonds with a system that used an animated embodied conversational agent (ECA) to deliver healthcare information than when a less realistic ECA or text was used to present the information. Such a system might easily be programmed to tailor the affective nature of response to specific types of user health query (rather than specific users).

This model is slightly at odds with the notion that designing for *user experience* provides a broader and more profitable explanation of requirements for system development than the more traditional notion of *usability* (see Section 3). We, therefore, suggest that *user experience* should be regarded, not as an alternative to affective computing, but as a broader, all encompassing term, in much the same way that it has been argued that 'affect' relates to 'emotion' (see Figure 4). This provides a consistency in the way that the term is applied.

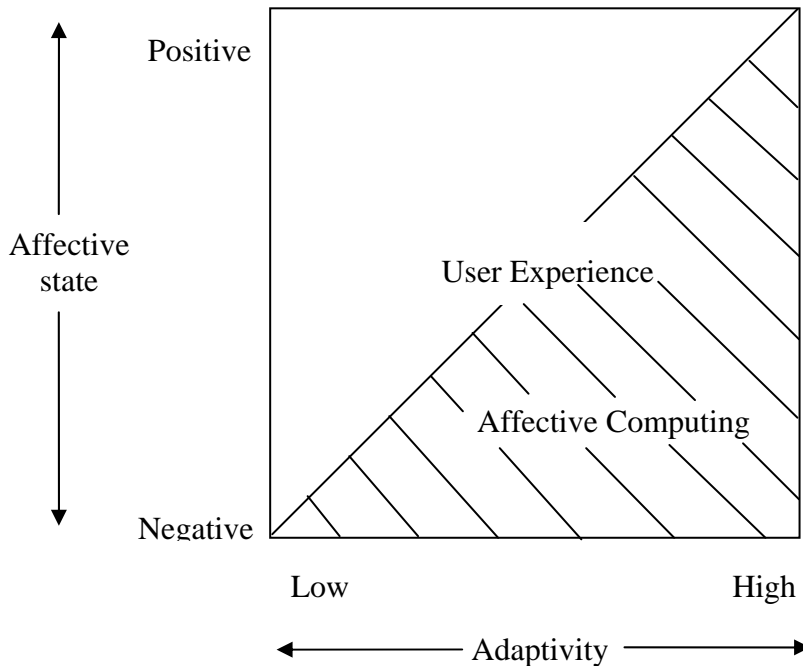


Figure 4. The association between ‘affective computing’ and ‘user experience’ (UX).

According to Hassenzahl and Tractinsky (2006) affective computing systems also tend to be concerned with negative (as opposed to positive) affective states (e.g., reducing user frustration). They contrast this with UX that focuses more strongly on achieving positive affective states (e.g., fun). Although some exceptions can be found (see Kapoor et al., 2004), this may be a reasonable description of the current position (see Section 5). However, it should perhaps be regarded as a bias, not an absolute. In the longer term there may be practical reasons that serve to sustain this position. Negative affective states are easier to detect (Cacioppo et al., 1990; Reeves & Nass, 2002). Therefore, systems that require detection of affect may be more effective when dealing with negative emotion.

4.2 Transmission of emotion-related information

Emotion-related computing system can be described in terms of the transmission and reception of emotion-related information. The model shown in Figure 5 illustrates this and provides a basis for a discussion of key components of these systems. The following paragraphs are numbered to correspond to the numbered components of the model. Emotion modelling/simulation systems, in which there is no human system component (see e.g.,

Bellman, 2002; Martinez-Miranda & Aldea, 2005) are regarded as outside the scope of this analysis.

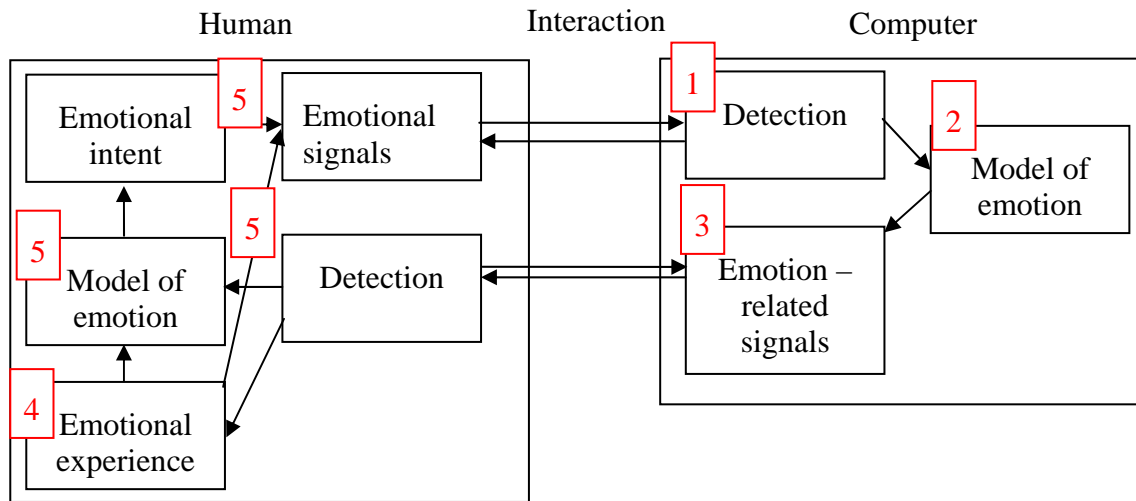


Figure 5. Transmission of emotional information between human and computer

1. This component of the model applies only to affective computing systems (i.e., it is not required for other types of UX system). If present it should be evaluated with reference to the criteria set out in Table 1 (Section 3.4). Particular consideration should be given to the reliability of emotion detection. Perfect detection is not possible, and does not happen in human-human interaction. An important question to be addressed is how standards of detection (a requirement specification for this system component) can be set. Human recognition of emotion provides one benchmark (see e.g., Picard et al., 2001). However, there are few studies that have tested the same stimuli using both computer and human assessors. This means that the magnitude of the emotion manipulation may be different making comparison difficult. When multi-modal assessment is being used it is important to be able to assess how much is gained by the inclusion of each channel and how much it ‘costs’ to be able to gather that data. For ‘real world’ systems, pragmatic solutions may not incorporate all possible channels. If a stable affective state is detected at this point in the system it may be because: i) there was no change in emotion to detect; ii) the method for detection was not sufficiently sensitive to that emotion; or iii) the expression of the emotion was suppressed by the user. Testing human as well as computer assessment (as above), and the use of multiple computer-based methods will go some way towards differentiating

alternatives i) and ii). Self-report methods of assessment (see Section 3.3.1) would be an appropriate means by which to address all alternatives.

2. An adaptive emotion-related system will require a model of emotion on which to base strategy selection, having aggregated (if more than one channel) the information from the detection stage. If the system is not adaptive then the model is embedded by the designer and relies on previous research evidence. The extent to which there are individual differences in users' responses to the system is a key issue and should be part of the empirical assessment process.
3. Production of emotion-related signals that will be transmitted to the user. These signals can take the form of messages (e.g., Klein et al., 2002) or can be alterations to the interface/interaction process. In all UX (including affective computing) systems the computer can transmit emotion-related information. We refer to this information as emotion-related so as not to imply that the computer is simulating emotion. In the case of most UX and affective computing systems the computer's actions are designed to manipulate the emotional state of the user. If this component of the model is interactive questions to be addressed are:
 - a. Will adaptivity confuse the user?
 - b. Will the user understand (or need to know) the basic rules by which adaption takes place?

In all circumstances we can ask:

- c. Will the signals be understood by the user?
- d. Does the user need to be aware (consciously) of the signals for benefit to accrue?

An important consideration for adaptive systems is whether the signals are well synchronised with the activities of the user. It is possible for signals of user emotion to be tested in laboratory experiments, found to be sensitive to differences between interface or task manipulations that were designed to alter user affect, but not with a temporal resolution that would enable them to be used as a basis for adapting the interface, i.e., interaction may proceed too quickly for adaptivity.

4. Obviously, the emotional experience of the user will depend on the nature of their experiences with the system. The implications of this, when interacting with an (adaptive) affective computing system, may be to reduce the within and between-user variability in affective state and therefore the diversity of affective experience. This prediction is made on the basis that an affective computing system goal would be the avoidance of certain affective states in the user (e.g., frustration) and the attainment of others (e.g., engagement). Given that the system has an effective model of emotion and generates appropriate emotion-related signals, the result would be interaction, and affective experience, that is more uniform.

However, this assumes an effective system. An important consideration, at this point in the model, is the extent to which individuals differ in their response to the same emotion-related signals. High variability would indicate that the computer must generate an emotion model for each individual. Low variability would indicate that a generic model may suffice. These would be important evaluations to make as part of system testing.

Emotional experience may depend on the context of use. For example, it would be of interest to know whether responses changed when the computer was being used to facilitate communication with a human, i.e., computer mediated communication (CMC) (although see Reeves & Nass, 2002). In these circumstances the use of emoticons has been found to be useful. Rivera, Cooke, & Bauhs (1996) provided some participants with six emoticons in a simulated CMC scenario. Those participants that had access to emoticons tended to use them. The availability of emoticons also changed the pattern of communication, with more participants in the 'emoticon' condition directing communication to an apparently angry (simulated) group member. However, there was also a tendency for participants in this group (the difference just failed to reach statistical significance) to report *greater* dissatisfaction with the communication experience.

5. Some emotional expression is relatively involuntary (e.g., changes in skin conductivity), and some is under user control (e.g., many aspects of facial and verbal expression) and can be used for the purposes of communication. In the model we show two paths to illustrate this. One depicts a flow of information from 'emotional

experience' to 'emotional signals'. The second path includes the user's model of emotion. This provides a model that can be used to predict how the person or thing that is to be the recipient of the emotional communication is likely to respond. To the extent that this route is available, emotional signals can be selected to serve the user's goals. This will have implications for the training of system responses.

When communication is dominated by the involuntary path there are important ethical issues to be considered (see Reynolds & Picard, 2004) and user acceptance may become a problem. For example, some participants in a Wizard of Oz study by Axelrod and Hone (2005) reported trying not to display emotion, when they thought the computer would try to assess this. However, apparently they failed in this goal. The extent to which the user is aware of their assessed affective state will also determine the apparent predictability of the system (see Berridge & Winkielman, 2003, for evidence supporting unconscious emotions).

4.3 Evaluating costs and benefits

Cost-benefit analysis can be an important component of usability work. However, it is typically concerned with demonstrating the value of a proposed usability intervention to those people responsible for managing the software development process, i.e., presenting a comparison of the benefits that would accrue from a usability intervention with the costs would be incurred by this intervention (see e.g., Rouse & Boff, 1997). We argue here that, in the context of affective computing systems (see Section 4.1 for a definition), there is particular value in considering the costs and benefits associated with the development and deployment of the emotion-related system component(s). If the ratio is unfavourable they should be redesigned or excluded. This proposal implies that it is possible to consider emotion-related system components independently from other system components. It might be argued that this will be difficult or impossible in a well integrated system. In the case of many UX systems we agree that this is the case. It would not be easy to determine what un-affective design elements would be and so comparisons are difficult. However, for affective computing systems, those that adapt to the affective state of the user, this is not a view we subscribe to. Many affective computing system components are relatively modular and 'with' and 'without' comparisons are feasible (see Section 5 of this report for examples).

There are costs associated with the development and deployment of an affective computing system that would not be incurred for an equivalent 'non-affective' version of the system. These go beyond financial costs associated with system development. Factors such as the level of user acceptance should be considered. For example, users might find the methods of assessing affective state intrusive, either because they require a physical interface with the user (e.g., placement of electrodes) or because the system may access affective information that the user would prefer to keep secret (see above). Similarly, it would be possible for an emotion-related system to place additional cognitive costs on the user. For example, if users feel that the system is less predictable because it is adapting to their affective states there may be additional cognitive costs associated with tracking and/or identifying current system state (cf., Kramer, 1991).

Benefits accruing from affective computing systems could also take several forms. Avoiding negative affective states, such as frustration is often cited as a goal (e.g., Klein et al., 2002). However, achieving positive affective states, such as interest or engagement, can also be important design goals. Success might be indicated, for example, by prolonged periods of system use (e.g., Klein et al., 2002). Given the impact of emotion on cognition (e.g., Isen *et al.*, 1987), improved user performance might also be a beneficial outcome.

An evaluation of costs and benefits needs to be done in the context of typical system use. For example, it is important to consider how frequently emotion-related system changes (adaptive) will occur. Under normal circumstances, experience of strong, recognizable emotion may happen relatively infrequently (at least in computing terms) (see Oatley & Duncan, 1994). It may be, therefore, that the system adapts to the state of the user relatively infrequently. If the costs are persistent (e.g., requiring the user to be attached to the computer via electrodes) and the benefits occasional this may indicate an unacceptable balance (depending also on other factors such as the degrees of cost and benefit). Assessing whether costs are outweighed by the benefits is complicated by the fact that different indicators are often used for these variables (see Rouse & Boff, 1997) and expert judgement may be the only viable alternative. However, it is important to note that a decision based on cost-benefit evidence need not be one of 'all or nothing' for the affective system components. For example, it may be that a less complex affective component could be used.

In making cost-benefit judgements it is important to bear in mind that they are being made in the context of a 'moving playing field'. As technology advances some costs are being

reduced (financial and others). So, from a commercial position, cost-benefits assessments must be continuously revised, and from a research perspective, cost-benefit assessments should not be too constraining, as what is resource expensive and difficult to achieve today, may be resource cheap, and easy to achieve tomorrow (cf. Rouse & Boff, 1997).

In Section 3.2 we presented arguments that the ‘traditional’ approach to usability assessment is not sufficient and that this concept should be replaced by reference to evaluating *user experience*. We now add another piece to the jigsaw that is particularly important for affective computing systems. For these systems, user experience must be considered in the context of the value being offered, i.e., what is the cost-benefit ratio.

Ultimately, the success or otherwise of any affective computing system (one that adapts to the user) will depend on the conjunction of the following factors:

- i) the effectiveness of emotion detection (Component 1), expressed as a probability of correct and incorrect detections;
- ii) the magnitude of the negative consequences (costs) to the user if the emotion is detected incorrectly;
- iii) the magnitude of the positive consequences (benefits) to the user if the emotion is detected correctly;
- iv) the likely frequency of system intervention (based on the user’s emotion).

The value of the system will be a function of the frequency of emotion classification events combined with the probability of misclassification and the associated costs if this should occur, versus the probability of a successful classification and the associated benefits if this should occur. Given human sensitivity towards negative outcomes (Frijda, 1988) a system bias might be appropriate (against negative outcomes). This would be an interesting topic for experimental investigation.

5. A selective review of affective computing studies

In this section, to get a sense of the ‘state of the art’, we present a review of quantitative studies of affective computing, that have used psychophysiological techniques to assess the affective state of the user. These studies all used tasks that were considered relevant to performance of ‘day-to-day’ computer based tasks (including gaming) rather than simulations of interaction with complex systems (see Section 3.3.3). This seems appropriate when evaluating the position for discretionary use systems. In the following paragraphs we consider the methods of measurement, the experimental tasks, how emotion was manipulated, and what the outcomes were. We do not include studies that have employed psychometrics as a major focus. These are examined in Workpackage deliverable D9f.

Studies were initially located on the basis of a search of psychology and computer science journals and conference proceedings using combinations of the key words: *affective, emotion, computing, physiological*. However, a ‘snowballing’ approach proved as useful, identifying papers that were either cited by or had cited others of interest. In total 12 papers reporting 13 studies were included as meeting the criteria for acceptance. These are indicated in bold in the reference list. However, in this section reference is made to further studies that did not meet the criteria for acceptance, but that provide useful background information.

5.1 Methods used to assess emotion

A variety of psychophysiological measures were used, as follows:

- i) facial EMG recordings of the corrugator and/or zygomaticus (Branco et al., 2005; Hazlett, 2003; Mahlke et al., 2006; Mandryk et al., 2006a; Partala & Surakka, 2004);
- ii) facial EMG recoding of the jaw (Mandryk et al., 2006b; Mandryk et al., 2006c);
- iii) face tracking (Ward, 2004);
- iv) galvanic skin response (GSR) (Lin & Hu, 2005; Mahlke et al., 2006; Mandryk et al., 2006a; Mandryk et al., 2006b; Mandryk et al., 2006c; Prendinger et al., 2005; Scheirer et al., 2002; Vinayagmamoorthy et al., 2006; Ward, 2004; Ward & Marsden, 2003);

- v) heart rate (Lin & Hu, 2005; Mahlke et al., 2006; Mandryk et al., 2006a; Mandryk et al., 2006b; Mandryk et al., 2006c; Vinayagamoorthy et al., 2006; Ward & Marsden, 2003);
- vi) BVP (Lin & Hu, 2005; Schierer et al., 2002; Ward & Marsden, 2003); and,
- vii) measures of respiration (Mandryk et al., 2006b; Mandryk et al., 2006c; Puri et al., 2005; Vinayagamoorthy et al., 2006).

Eight of the 13 studies included one or more psychometric measures. Nine studies included more than one psychophysiological measure. There are potentially important differences between studies in the way some of the psychophysiological measures were recorded. In particular, from the reports it would appear that there is some variation in the approach to measuring baselines. It is interesting to note that no studies were located that used EEG as a measure of affective state, even though there is evidence for emotion-related hemispheric asymmetry of activation (Cacioppo et al., 2000; Reeves & Nass, 2002). Left hemisphere activation tends to be associated with positive affect and right hemisphere activity with negative affect. However, EEG is more intrusive for the user (although not necessarily the task) than some other measures of psychophysiology.

5.2 Types of task and experimental manipulations

The task used in six of the studies was a game and for a further two it was a quiz. One of the studies used a word processing task, three required information retrieval (one involved locating objects in VR), and one used a task related to mobile phone use.

A rather limited range of emotion-evoking situations have been considered in experimental studies. Three of the studies used a task difficulty manipulation, five used manipulations of usability (as per ISO 9241), two used manipulations of the affective content of an ECAs interaction with the user following system malfunctions. The experimental scenarios for six of the studies involved the system malfunctioning or being difficult in its interaction. Generally the studies involved placing the user in a negative affective state. The two ECA studies were concerned with how best to redeem that situation.

An important variable, and one that is difficult to assess is the strength of the manipulation of user affect (e.g., the amount of interface disruption that was introduced). In

some cases this did seem relatively extreme (e.g., use of very persistent pop-ups that obscured an important area of the screen. This leaves open the question as to whether some of these methods of assessment would be successful within the range of affect variation that might typically be experienced by users (cf. Oatley & Duncan, 1994). In other words, the strong manipulations used in some studies may not be representative of ‘normal’ computing events (see Ward, 2004).

5.3 Outcomes

All of the studies that assessed EMG corrugator activity found a significant difference in activity resulting from the manipulation of affect. Zygomaticus activity, however, in two of four studies in which it was assessed showed the opposite pattern to that predicted (i.e. there was greater activity in a negative affect condition). There are other examples of negative emotional states leading to elevated zygomaticus activity. Possible explanations include ‘distress smiling’ and cross-talk from other muscles (Cacioppo et al., 2000). When considering methodological differences between these studies (two finding zygomaticus activity associated with negative affect and two with positive affect) task and context factors offer possible explanations. Both of the studies that found an association between zygomaticus and positive affect had ‘social/communicative’ components embedded within the task. In one of these, participants played a game against a friend, a stranger, or the computer. It was hypothesised that social gaming would be more pleasurable for participants and results supported this hypothesis. In the other study, a simulated malfunction was followed either by no message, a negative message, or a positive message. Zygomaticus activity was greatest when participants received the positive message. It was also greater in the negative message condition than the no message condition. From this we can speculate that zygomaticus activity is associated with a range of eliciting factors, one of which is elevated task difficulty and another that concerns social interaction.

Galvanic Skin Response (GSR) was used in 10 of the 13 studies, and significant differences relating to the experimental manipulation of affect were found in eight (although for three of these GSR was contributing to a multi-method assessment). One of the studies that produced a non significant result used a particularly small sample (n=7). Generally this would seem to be a positive set of results for GSR. However, it should be remembered that GSR is indicative of arousal but not valence (Prendinger et al., 2005). To be able to draw firm conclusions from GSR data it must be included as part of a multi-method approach.

Heart rate (HR) was included in six studies and contributed to the identification of significant differences associated with affect manipulations in just two, and these were both multi-method studies in which data was aggregated. As mentioned earlier, in the Human Factors literature, HRV seems to be a more effective measure.

A multi-method approach to assessment may provide greater reliability of measurement (Picard & Bryant, 2005). There are two possible reasons for this. First, it may be that each measure provides a rather ‘noisy’ assessment of the affective state in question. On the basis of psychometric principles (see deliverable D9f), given that a ‘true score’ exists for that affective state, aggregation of many ‘noisy’ measures will give a better approximation of the true score. Alternatively (or perhaps in addition), it may be that different methods of measurement ‘tap into’ different aspects of affective state. By combining the assessments taken from multiple methods it is possible to assess affect in the context of a multi-dimensional patterning of affective states (Scherer, 2005). Consistent with this latter position, correlations between different methods of assessing affect can be rather weak (e.g., Mahlke, Minge, & Thüring, 2006). In support of the multi-modal approach, three of the reviewed studies were of this type and all three found significant differences between experimental conditions.

5.4 Effects of task difficulty: Old wine in new bottles?

A potential problem with the data reported in the studies reviewed is that differences between conditions, in many instances, may relate to differences in the effort expended by the participant as a result of differences in task demands. This may be the case even when demand was not explicitly altered. Consider, for example, studies by Mandryk, Atkins, & Inkpen (2006) and Mandryk, Inkpen, & Calvert (2006). The experimental manipulation involved participants playing against another person rather than the computer, and this was hypothesised to influence affect. However, it might be argued that participants were trying harder, they were making more effort, when playing against another person and that this is the cause of differences in psychophysiological measures.

From this position it can be argued that much, if not most, of the available studies on affective computing are replicating results obtained by Human Factors researchers over the

last 30 years or so, when examining operator workload. It is important that this issue is addressed in the design of future affective computing studies.

5.5 Differentiating emotions

Can these methods identify different qualities of affect? The study by Ward (2004) included experimental manipulations designed to evoke: i) amusement; and ii) surprise. Face tracking was less sensitive to the former. Only 5 out of 15 participants were judged to have greater post event facial movement in this condition compared with 15 out of 15 for the surprise event. However, this may be nothing to do with the qualitative difference between emotions it may simply reflect differences in the scale of experienced emotion, with the manipulation of amusement not as strong as the manipulation of surprise. Multi-modal studies, in which data is obtained from different psychophysiological and behavioural sources (e.g., Mandryk et al., 2006) may provide a useful way forward, in this regard. Differentiation would be supported by patterns of activity elicited by different emotions. In addition, self report measures provide a useful check of results derived from psychophysiology.

6. Conclusions

This report presents an examination of issues relating to the assessment of the usability of emotion-related systems. The emphasis is on quantitative assessments and in this respect the report is complementary to other activities in WP9 of the HUMAINE NoE. In this final section we distil the material covered into a series of brief summaries and recommendations for action.

6.1 ‘Usability Testing’: An appropriate methodology for the evaluation of emotion-related systems?

In this report it is suggested that the approach to ‘usability testing’, that has been so dominant in researchers’ and practitioners’ thinking about the software development process over recent years, is too constrained for the purpose of evaluating emotion-related systems. There is a growing realisation amongst researchers in the area that a more comprehensive treatment of user motivation is required (Dillon, 2001; Hornbaek, 2006). However, the current research literature provides little in the way of guidance as to how this transition should be made. Dillon (2001) and Hornbaek (2006) propose that the traditional framework for assessment, with its focus on effectiveness, efficiency, and satisfaction (ISO 9241) should be extended to provide broader conceptualisations of the *process* and *outcome* of interaction, in addition to consideration of the *affective experience* of the user. The combination of assessments of process with those of affect appears promising, although there are important practical issues, such as the need for assessment methods that have the appropriate temporal resolution to permit this sequential form of analysis.

An analysis of costs and benefits (financial and other) associated with affective computing components was advocated in Section 4 as part of a global system assessment. It was argued that adaptive systems of this type are sufficiently modular in design to permit this. However, it should be recognised that with advances in technology system costs will tend to decline over time.

6.2 Methods of assessment

Existing methods of data gathering were reviewed in Section 3 and their suitability for evaluating emotion-related computing systems considered. In Section 5 a selective review of affective computing studies that have used psychophysiological techniques was reported. We have some concerns over limitations in the types of experimental manipulations that have been applied in many studies. These are discussed in Section 6.3. However, generally it would seem that many existing ‘usability’ data gathering methods can readily be applied in the broader domain of affective computing. Self-report methods have well known limitations but always provide a useful ‘touchstone’. A good deal of relevant research activity has taken place, using self-report methods. This is described in HUMAINE project deliverable D9f. Some new methods (at least from a ‘usability’ perspective) will need to be added for the assessment of non-verbal behaviour, e.g., gesture, posture. These are potentially important clues as to the user’s affective state, and are not well represented by the current battery of methods.

Psychophysiological assessments have been used extensively for a number of years in the area of Human Factors for the appraisal of operator workload. In recent years these methods are beginning to be used for the assessment of affective computing applications. Although, as will be discussed, there are limitations to the conclusions that can be drawn on the basis of some of the evidence produced, some psychophysiological methods are fairly consistently detecting differences between conditions that have been designed to manipulate the affective state of the user. Galvanic Skin Response (GSR) was an effective method in this respect in the studies reviewed (see Section 5). Although if this method is used in isolation the valence of the users response to an experimental manipulation (e.g., different levels of task difficulty) can be ambiguous. Corrugator muscle activity also appears to be a reliable predictor of condition differences. However, zygomaticus muscle activity was inconsistent across the studies reviewed. It was suggested that factors relating to social responding may be the cause (see Section 5 for details).

Several of the reviewed studies assessed heart rate, but none found it predictive of experimental conditions, unless it was included as part of a multi-modal approach. In the Human Factors literature relating to workload assessment Heart Rate Variability seems to prove the more reliable index, and it might also be useful in this context.

Heuristic review was identified as an existing usability method that could be revised to better suit application to emotion-related systems. Heuristic guidelines have been developed for other specific areas of application, such as virtual reality applications and ambient displays. A similar approach could be taken for affective computing systems. This would have the added benefit of making key design principles explicit, and promoting discussion/evaluation of these.

Surprisingly we found no studies of affective computing that used EEG recording. Given that EEG has been reported to be sensitive to differences in hemispheric activity that arise from differences in the valence of affective state (see Cacioppo et al., 2000), it would seem a promising method for this domain of application, although a potentially intrusive/inconvenient one for the user.

6.3 The manipulation of affect

In the selective review of studies reported in Section 5, we felt that the nature of the experimental manipulations used was not always convincing and leaves some results open to alternative explanation. In many instances the manipulation was either explicitly or may have implicitly been one of task demand. This does not equate with testing negatively valenced states that are the target of much thinking in the area of affective computing (e.g., frustration). This line of argument is expanded in Section 5. However, if valid, and the key difference between experimental conditions was cognitive load, the results of many of the psychophysiological assessments do not contribute much beyond those of the large number of Human Factors studies of operator workload that have been conducted over the last 25 years. It is suggested that future studies should focus on precise affective manipulations that target key affective states and that can be validated.

6.4 Standardisation

Also from the review of studies, presented in Section 5, that have manipulated and assessed user affect it would seem that a means of standardizing manipulation strength is required. It is not clear how the studies compare, one to another, in this respect, or whether the manipulations being used in experimental studies of emotion-related computing systems are of an appropriate scale (i.e. are realistic). This makes interpretation of usability problematic. When assessing the value of designing for user experience it is necessary to

know that effects obtained are of a similar magnitude that would be obtained in 'real world' situations. This difficulty might be overcome by gathering standardization data that references everyday experiences of emotion (e.g., Oatley & Duncan, 1994).

A related measurement issue concerns the effectiveness with which the computer is able to judge the affective state of the user on the basis of psychophysiological assessments (regardless of the size of the manipulation). The use of human judges of the same information would provide a standard against which this could be readily be assessed. In most cases the current dominant paradigm for assessing emotion recognition, involving the classification of emotion, would need to be replaced with one that is compatible with a dimensional model. However, this does not seem an insurmountable hurdle.

Finally, with regard to standardization issues, assessments of the magnitude of individual differences are important. An important issue is whether affective computing systems can be effective with a single model of users' responses, which provides a basis for the system adapting, or whether separate models are required for each user.

6.5 Differentiating emotions

Hanssenszahl & Tractinsky (2006) suggest that affective computing systems are predominantly concerned with negative emotions. When considering existing research there does seem to be a strong bias in this direction, with many studies focusing on user frustration or conditions of high stress/demand. In this context, relatively little work has been done on detecting and adapting to positive emotions, e.g., engagement with, or interest in, an interface. This may be due to a bias in the capacity of psychophysiological methods to detect emotion, such that negative emotions are more easily detected (Cacioppo, 2000), it may be because researchers still closely align themselves with 'traditional' models of system usability in which they look for problems to solve rather focusing on opportunities, or it may be due to a human bias to regard things with negative valence as more significant (Frijda, 1988).

Information retrieval seems a particularly promising task domain for investigating user experiences of engagement. The development of computer systems and the WWW makes information retrieval a routine task for many people, but one that is increasingly complicated by the volume of electronic data. Being able to determine the relevance of specific items of information for the user, i.e., the extent to which users are interested in or engage with an item

of information, is a very important goal. Empirical support for this task domain would have the potential for substantial practical advantage.

A final question to be addressed concerns the extent to which differentiation of emotional states is required by affective computing systems. The circumplex model from Scherer (2005), included in Section 3, suggests that at least two dimensional representation is required to be able to differentiate some of the affective states that are of most interest to the HCI research community. On the basis of existing evidence from affective computing studies, it is not clear that this can be achieved from psychophysiological measurement alone (cf. Cacioppo et al., 2000). A multi-modal approach is suggested as the most promising, in which participants responses are examined for patterns of response across the different measures. Inclusion of ‘process’ variables, such as keystroke data would make for rich and interesting analyses. The costs and benefits of greater/lesser differentiation should be appraised.

6.6 Final remarks

In this report we have reviewed the ‘traditional’ concept of usability testing and examined its application to the field of emotion-related computing. To facilitate this we have reviewed literature, including a set of empirical studies on the topic of affective computing. This has led to the conclusions that the concept of usability should be replaced by a broader construct that incorporates and supplements the existing methods and approaches. On the basis of our review we have begun to describe some of these revisions.

References

- Allanson, J. & Fairclough, S.H. (2004). A research agenda for physiological computing. Interacting with Computers, 16, 857-878.
- Anttonen, J. & Surakka, V. (2005). Emotions and heart rate while sitting on a chair. In Proceedings of CHI 2005. NY: ACM. Pp. 491-499.
- Axelrod & Hone (2005). E-motional advantage: Performance and Satisfaction gains with affective computing. In Proceedings of CHI 2005. ACM.
- Baeza-Yates, R. & Ribero-Neto, B. (1999). Modern Information Retrieval. Harlow: Pearson.
- Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology, 70, 614–636.
- Barrett, L.F. & Wager, T.D. (2006). The structure of emotion: Evidence from neuroimaging studies. Current Directions in Psychological Science, 15, 79-83.
- Bellman, K.L. (2002). Emotions: Meaningful mappings between the individual and its world. In R. Trasppl, P. Petta, & S. Payr (Eds.), Emotions in Humans and Artefacts. Cambridge, MA: MIT Press.
- Bently, T., Johnston, L., von Baggo, K. (2005). Evaluation using cued-recall debrief to elicit information about users' affective experiences. In Proceedings of OZCHI 2005, Canberra, Australia.
- Berridge, K.C. & Winkielman, P. (2003). What is an unconscious emotion? (The case for unconscious "liking"). Cognition and Emotion, 17, 181-211.
- Bessi re, K., Newhagen, J.E., Robinson, J.P., Shneiderman, B. (2006). A model for computer frustration: The role of instrumental and dispositional factors on incident, session, and post-session frustration and mood. Computers in Human Behavior, 22, 941-961.
- Bevan, N. (2006). Practical issues in usability measurement. Interactions, 6, 42-43.
- Bickmore, T. & Mauer, D. (2006). Modalities for building relationships with handheld computer agents. In Proceedings of CHI 2006. NY: ACM. Pp. 544-549.
- Branco, P., Firth, P., Encarna o, M., & Bonato, P. (2005). Faces of emotion in human-computer interaction. In CHI 2005, April 2-7, Portland, Oregon. ACM.**
- Busso, C. Deng, Z., Yildirim, S., Bulut, M., Lee, C.M., Kazemazdeh, A., Lee, S., Neumann, U., & Narayanan, S. (2004). Analysis of emotion recognition using facial expressions, speech and multimodal information. In Proceedings of ICMI'04, Oct 13-15, State College Pennsylvania. ACM.

- Cacioppo, J.T., Bertson, G.C., Larsen, J.T., Poehlmann, K.M., & Ito, T.A. (2000). The psychophysiology of emotion. In M. Lewis & J.M. Haviland-Jones (Eds.), Handbook of Emotions. 2nd Edition. NY: Guilford Press.
- Carver, C.S. (2004). Negative affects deriving from the Behavioural Approach System. Emotion, 4, 3-22
- Cheung, C.M.K. & Lee, M.K.O. (2005). The asymmetric effect of website attribute performance on satisfaction: An empirical study. In Proceedings of the 38th Hawaii International Conference on System Sciences. IEEE.
- Christie, I.C. & Friedman, B.H. (2004). Autonomic specificity of discrete emotion and dimensions of affective space: A multivariate approach. International Journal of Psychophysiology, 51, 143-153.
- Cockton, G. (2004). Doing to be: Multiple routes to affective interaction. Interacting with Computers, 16, 683-691.
- Conati, C. (2004). How to evaluate models of user affect? Lecture Notes in Computer Science, 3068, 288-300.
- Crosby, M.E., Auernheimer, B., Aschmwenden, C., & Ikehara, C. (2001). Physiological data feedback for application in distance education. In Proceedings of PUI 2001. NY: ACM. Pp. 1-5.
- Desmet, P.M.A. (2003). Measuring emotions: Development and application of an instrument to measure emotional responses to products. In M.A. Blythe, A.F. Monk, K. Overbeeke, & P.C. Wright (Eds.), Funology: From Usability to Enjoyment. Dordrecht: Kluwer Academic Publishers.
- Desurvire, H., Caplan, M. & Toth, J.A. (2004). Using heuristics to evaluate the playability of games. In Ext. Abst. CHI 2004. NY: ACM Press. Pp. 1509-1512.
- Diaper, D. (1990). Task Analysis for Human-Computer Interaction. Upper Saddle River, NJ: Prentice Hall.
- Dillon, A. (2001). Beyond usability: Process, outcome, and affect in human-computer interactions. The Canadian Journal of Information and Library Science, 26, 4, 57-69.
- Ekman, P. (1977). Biological and cultural contributions to body and facial movement. In J. Blacking (ed.). Anthropology of the Body, London: Academic Press, pp. 34-84.
- Ekman, P. (1999). Basic Emotions. In T. Dalgleish and T. Power (Eds.) The Handbook of Cognition and Emotion. Sussex, U.K.: John Wiley & Sons, Ltd.
- Frijda, N.H. (1986). The Emotions. Cambridge, England: Cambridge University Press.
- Frijda, N.H. (1988). The laws of emotion. American Psychologist, 43, 349-358.

- Frøkjær, E., Hertzum, M., & Hornbæk, K. (2000). Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated? CHI Letters, *2*, 1, 345-352.
- Gedgia, G., Hamborg, K.C., & Duntsch, I. (1999). The IsoMetrics inventory: An operationalization of ISO 9241-10 supporting summative and formative evaluation of software systems. Behaviour & Information Technology, *18*, 151-164.
- Gilleade, K.M. & Dix, A. (2004). Using frustration in the design of adaptive videogames. In Proceedings of ACE 2004. NY: ACM. Pp. 228-232.
- Goleman, D. (1995). Emotional Intelligence. New York: Bantam Books.
- Gulliksen, J. Boivie, I., Persson, J. Hektor, A., & Herulf, L. (2004). Making a difference – a survey of the usability profession in Sweden. In Proceedings of NordiCHI '04. NY: ACM. Pp. 207-215.
- Hassenzahl, M., Platz, A., Burmester, M., & Lerner, K. (2000). Hedonic and ergonomic quality aspect determine a software's appeal. CHI Letters, *2*, 1, 201-208.
- Hassenzahl, M. & Tractinsky, N. (2006). User experience – a research agenda. Behavior & Information Technology, *25*, 91-97.
- Hazlett, R. (2003). Measurement of user frustration: A biologic approach. In CHI 2003, April 5-10, Ft. Lauderdale, Florida. ACM.**
- Henderson, R., Podd, J., Smith, M., & Varela-Alvarez, H. (1995). An examination of four user-based software evaluation methods. Interacting with Computers, *7*, 412-432.
- Hornbaek, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. International Journal of Man-Machine Studies, *64*, 79-102.
- Hudlicka, E. (2003). To feel or not to feel: The role of affect in human-computer interaction. International Journal of Human-Computer Studies, *59*, 1-32.
- Iqbal, S.T., Adamczyk, P.D., Zheng, X.S., & Bailey, B.P. (2005). Toward an index of opportunity: Understanding changes in mental workload during task execution. In Proceedings of CHI 2005. New York: ACM. Pp. 311-320.
- ISO 9241-11. (1998). Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs). Part 11: Guidance on Usability. Geneva: International Organization for Standardization
- Isen, A.M. (1985). The asymmetry of happiness and sadness in effects on memory in normal college students. Journal of Experimental Psychology: General, *114*, 388-391.
- Isen, A.M. (2004). Some perspectives on positive feelings and emotions. In A.S.R. Manstead, N. Frijda, & A. Fischer (Eds.), Feelings and Emotions: The Amsterdam Symposium. Cambridge: Cambridge University Press.

- Isen, A.M. & Daubman, K.A. (1984). The influence of affect on categorization. Journal of Personality and Social Psychology, 47, 1206-1217.
- Isen, A.M., Daubman, K.A & Nowicki, G.P. (1987). Positive affect facilitates creative problem solving. Journal of Personality and Social Psychology. Vol 52(6), 1122-1131.
- Isaacowitz, D.M. (2006). Motivated gaze: The view from the gazer. Current Directions in Psychological Science, 15, 68-72.
- Isaacowitz, D.M., Wadlinger, H.A., Goren, D., & Wilson, H.R. (2006). Is there an age-related positivity effect in visual attention? A comparison of two methodologies. Emotion, 6, 511-516.
- Ivory, M.Y. & Hearst, M.A. (2001). The state of the art in automating usability evaluation of user interfaces. ACM Computing Surveys, 33, 470-516.
- Izard, C.E. (1971). The face of emotion. East Norwalk, CT, US: Appleton-Century-Crofts.
- Izard, C. E. (1977). Human Emotions. New York: Plenum Press.
- Joachims, T., Granka, L., Pan, B., & Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. In Proceedings of SIGIR'05. NY: ACM. Pp. 154-161.
- Johnson, D. & Wiles, J. (2003). Effective affective user interface design in games. Ergonomics, 46, 1332-1345.
- Jordan, P.W. (2000). Designing Pleasurable Products. London: Taylor & Francis.
- Kapoor, A., Picard, R.W., & Ivanov, Y. (2004). Probabilistic combination of multiple modalities to detect interest. In Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04). IEEE.
- Karat, J. (1997). User-centred software evaluation methodologies. In M. Helander, T.K. Landauer, & P. Prabhu (Eds.), Handbook of Human-Computer Interaction. 2nd Edition. Amsterdam: Elsevier Science.
- Kirakowski, J. & Corbett, M. (1993). SUMI: The software usability measurement inventory. British Journal of Educational Technology, 24, 210-212.
- Klein, J., Moon, Y., & Picard, R.W. (2002). This computer responds to user frustration: Theory, design, and results. Interacting with Computers, 14, 119-140.**
- Kramer, A.F. (1991). Physiological metrics of mental workload: A review of recent progress. In D.L. Damos (Ed.), Multiple-Task Performance. London: Taylor & Francis.
- Lang, P.J., Bradley, M.M. & Cuthbert, B.N. (1997). Motivated attention: Affect, activation, and action. In P.J. Lang, R.F. Simmons, & M. Balaban (Eds). Attention and orienting: Sensory and Motivational Processes. Mahwah, N.J.: Erlbaum.

- Law, E.L-C. & Hvannberg, E.T. (2004). Analysis of strategies for improving and estimating the effectiveness of heuristic evaluation. In Proceedings of NordiCHI'04. New York: ACM. Pp. 241-250.
- Lazarus, R.S. (1982). Thoughts on the relations between emotion and cognition. American Psychologist, 37, 1019-1024.
- Lin, T., Hu, W., Omata, M., & Imamiya, A. (2005). Do physiological data relate to traditional usability indexes. Proceedings of OZCHI 2005. New York: ACM.**
- Lin, T. & Imamiya, A. (2006). Evaluating usability based on multimodal information: An empirical study. In Proceedings of ICMI'06. NY: ACM. Pp. 364-371.
- Lopes, P.N., Salovey, P., Cote, S., Beers, M. & Petty, R.E. (2005). Emotion regulation abilities and the quality of social interaction. Emotion, 5 (1) 113-118.
- Malke, S., Minge, M., & Thüring, M. (2006). Measuring multiple components of emotions in interactive contexts. In Proceedings of CHI 2006. Montreal, Quebec, Canada. ACM.**
- Mandryk, R.L., Atkins, M.S., & Inkpen, K.M. (2006). A continuous and objective evaluation of emotional experience with interactive play environments. In Proceedings of CHI 2006. NY: ACM**
- Mandryk, R.L., Inkpen, K.M., & Calvert, T.W. (2006). Using psychophysiological techniques to measure user experience with entertainment technologies. Behaviour & Information Technology, 25, 2, 141-158.**
- Mankoff, J., Dey, A.K., Hsieh, G., Kientz, J., Lederer, S., & Ames, M. (2003). Heuristic evaluation of ambient displays. In Proceedings of CHI 2003, 5, 1, 169-176.
- Mao, J-Y. Vrendenburg, K., Smith, P.W., & Carey, T. (2005). The state of user-centred design practice. Communications of the ACM, 48, 3, 105-109.
- Martinez-Miranda, J. & Aldea, A. (2005). Emotions in human and artificial intelligence. Computers in Human Behavior, 21, 323-341.
- McNeese, M.D. (2003). New visions of human-computer interaction: Making affect compute. International Journal of Human-Computer Interaction, 59, 33-53.
- Nielsen, J. (1993) Usability Engineering. London: Academic Press.
- Nielsen, J. (1992). Finding usability problems through heuristic evaluation. In Proceedings of CHI '92. New York: ACM. Pp. 373-380.
- Nielsen, J. & Molich, R. (1990). Heuristic evaluation of user interfaces. In Proceedings of CHI '90. New York: ACM. Pp. 249-256.
- Nisbett, R.E. & Wilson, T.D. (1977). Telling more than we can know: Verbal reports on mental processes. Psychological Review, 84, 231-259.

- Norman, D.A. (1988) *The Psychology of Everyday Things*. New York: Basic Books.
- Norman, D.A. (2004). Emotional Design. New York: Basic Books.
- Oatley, K. & Duncan, E. (1994). The experience of emotions in everyday life. Cognition & Emotion, 8, 369-381.
- O'Donnell, R.D. & Eggemeier, E.T. (1986). Workload assessment methodology. In K.R. Boff, L. Kaufman, & J.P. Thomas (Eds.), Handbook of Perception and Human Performance. Vol. II. NY: Wiley.
- Okada, H. & Asahi, T. (1999). GUITESTER: A log-based usability testing tool for graphical user interfaces. IEICE Transactions on Information and Systems, E82D, 6, 1030-1041.
- Parkinson, B. (2005). Do facial movements express emotions or communicate motives? Personality and Social Psychology Review, 9, 278-311.
- Partala, T. & Surakka, V. (2004). The effects of affective interventions in human-computer interaction. Interacting with Computers, 16, 295-309.**
- Picard, R.W. (1997). Affective Computing. Cambridge, MA: MIT Press.
- Picard, R. & Daily, S.B. (2005). Evaluating affective intentions: Alternatives to asking what users feel. Conference on Human Factors in Computing Systems. Workshop on Innovative Approaches to Evaluating Affective Interfaces. Portland, OR, April 2-7.
- Picard, R.W., Vyas, E., Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23, 1175-1191.
- Plutchik, R. (1962). The Emotions: Facts, Theories, and a New Model. New York: Random House.
- Plutchik, R. (1980). Emotion: A Psychoevolutionary Synthesis. New York: Harper & Row.
- Power, M.J. (2006). The structure of emotion: An empirical comparison of six models. Cognition and Emotion, 20, 5, 694-713.
- Preece, J., Rogers, Y. and Sharp, H. (2002) *Interaction Design*. New York: Wiley.
- Prendinger, H., Ma, C., Yingzi, J., Nakasone, A. & Ishizuka, M. (2005). Understanding the effect of life-like interface agents through users' eye movements. In Proceedings of ICMI'05. NY: ACM. Pp. 108-115.
- Prendinger, H., Mori, J., Ishizuka, M. (2005). Using physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game. International Journal of Human-Computer Studies, 62, 231-245.**
- Puri, C., Olson, L., Pavlidis, I., Levine, J., & Starren, J. (2005). StressCam: Non-contact measurement of users' emotional states through thermal imaging. In CHI2005, Late Breaking Results: Posters. ACM. pp. 1725-1728.

- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. Psychological Bulletin, 124, 372-422.
- Reeves, B. & Nass, C. (1996). The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. Stanford, CA: CSLI Publications.
- Reynolds, C. & Picard, R. (2004). Affect sensors, privacy, and ethical contracts. In Proceedings of CHI2004, pp. 1103-1106.
- Rivera, K., Cooke, N.J., & Bauhs, J.A. (1996). The effects of emotional icons on remote communication. In CHI '96. NY: ACM Press. Pp. 99-100.
- Rouse, W.B. & Boff, K.R. (1997). Assessing cost/benefits of human factors. In G. Salvendy Ed.), Handbook of Human Factors and Ergonomics, 2nd Edition. New York: Wiley.
- Russell, J.A. (2003). Core affect and the psychological construction of emotion. Psychological Review, 110, 145-172.
- Sauro, J. & Kindlund, E. (2005). A method to standardize usability metrics into a single score. In Proceedings of CHI 2005. NY: ACM. Pp. 401-409.
- Scheirer, J., Fernandez, R., & Picard, R.W. (1999). Expression glasses: A wearable device for facial expression recognition. In Proceedings of CHI'99. NY: ACM. Pp. 262-263.
- Scheirer, J., Fernandez, R., Klein, J., & Picard, R.W. (2002). Frustrating the user on purpose: A step toward building an affective computer. Interacting with Computers, 14, 93-118.**
- Scherer, K. R. (2005). What are emotions? And how can they be measured? Social Science Information, 44, 4, 693-727.
- Selker, T. (2004). Visual attentive interfaces. BT Technology Journal, 22, 146-150.
- Shneiderman, B. (1998) Designing the User Interface: Strategies for Effective Human-Computer Interaction. 3rd edition. Reading, MA: Addison Wesley Longman.
- Smith, P.A. (1996). Toward a practical measure of hypertext usability. Interacting with Computers, 8, 365-381.
- Smith, S.L. & Mosier, J.N. (1986). Design Guidelines for Designing User Interface Software. Technical Report MTR-10090, The MITRE Corporation, Bedford, MA.
- Sutcliffe, A. & Gault, B. (2004). Heuristic evaluation of virtual reality application. Interacting with Computers, 16, 831-849.
- Sweetsner, P & Wyeth, P. (2005). GameFlow: A model for evaluating player enjoyment in games. ACM Computers in Entertainment, 3, 3, Article 3A.

- Vinayagmoorthy, V., Brogni, A., Steed, A., & Slater, M. (2006). The role of posture in the communication of affect in an immersive virtual environment. In Proceedings of VRCIA, ACM. Pp. 229-236.**
- Virzi, R.A. (1997). Usability inspection methods. In M. Helander, T.K. Landauer, & P. Prabhu (Eds.), Handbook of Human-Computer Interaction, 2nd Edition. North Holland: Elsevier Science.
- Vrendenburg, K., Mao, J-Y., Smith, P.W., & Carey, T. (2002). A survey of user-centred design practice. CHI Letters, 4, 1, 471-478.
- Ward, R. (2004). An analysis of facial movement tracking in ordinary human-computer interaction. Interacting with Computers, 16, 879-896.**
- Ward, R.D. & Marsden, P.H. (2003). Physiological responses to different web page designs. International Journal of Human-Computer Studies, 59, 199-213.**
- Ward, R.D. & Marsden, P.H. (2004). Affective computing: problems, reactions and intentions. Interacting with Computers, 16 (4), 707-713.
- Watson, D. & Tellegen, A. (1985). Toward a consensual structure of mood. Psychological Bulletin, 98, 219-235.
- Watson, D., Wiese, D., Vaidya, J., & Tellegen, A. (1999). The two general activation systems of affect: Structural findings, evolutionary considerations, and psychobiological evidence. Journal of Personality and Social Psychology, 77, 863-877.
- Wensveen, S., Overbeeke, K., & Djajadiningrat, T. (2002). Push me, shove me and I show you how I feel: Recognising mood from emotionally rich interaction. In Proceedings of DIS2002, NY: ACM. Pp. 335-340.
- Wilson, G.M. (2001). Psychophysiological indicators of the impact of media quality on users. In Proceedings of CHI 2001. Doctoral Consortium. ACM. Pp. 95-96.
- Wilson, G.M. & Sasse, M.A. (2000). The head or the heart? Measuring the impact of media quality. In Proceedings of CHI 2000. Interactive Posters. ACM. Pp. 117-118.
- Wixon, D. & Wilson, C. (1997). The usability engineering framework for product design and evaluation. In M. Helander, T.K. Landauer, & P. Prabhu (Eds.), Handbook of Human-Computer Interaction. 2nd Edition. North Holland: Elsevier Science.
- Zhang, P. & von Drang, G.M. (2000). Satisfiers and dissatisfiers: A two-factor model for website design and evaluation. Journal of the American Society for Information Science, 51, 14, 1253-1268.