

# humaine

**D9f**

**Usability Testing Emotion-Oriented Computing Systems:  
Psychometric Assessment**

*Workpackage 9 Deliverable*



**Date: 30<sup>th</sup> September 2006**

<b>IST project contract no.</b>	507422
<b>Project title</b>	<b>HUMAINE</b> <b>Human-Machine Interaction Network on Emotions</b>
<b>Contractual date of delivery</b>	<i>September 30, 2006</i>
<b>Actual date of delivery</b>	<i>January 25, 2006</i>
<b>Deliverable number</b>	D9f
<b>Deliverable title</b>	Usability Testing Emotion-Oriented Computing Systems: Psychometric Assessment
<b>Type</b>	Report
<b>Number of pages</b>	51
<b>WP contributing to the deliverable</b>	WP 9
<b>Task leader</b>	UNIVLEEDS
<b>Author(s)</b>	S.J. Westerman, P.H. Gardner, E.J. Sutherland
<b>EC Project Officer</b>	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 in HUMAINE	5
1.1	The field covered by Work Package 9	5
1.2	The research objectives	5
1.3	Main elements of the deliverable	6
2	Setting the scene	8
2.1	Emotion-related computing	8
2.2	The characteristics of emotion	8
2.3	The nature of psychometric tests	9
2.4	Usability testing	12
3.0	Computing specific psychometric tests of emotion/affect	14
3.1	Satisfaction: Affect or attitude?	14
3.2	Standardising the measurement of user satisfaction	15
3.3	Satisfaction as a psychological/design tool	16
3.4	Extending the measurement of user affect	17
3.4.1	Computer anxiety	19
3.4.2	Flow, fun, and playfulness	19
3.4.3	Trust and loyalty	20
3.4.4	Frustration	22

---

Section	Title	Page
3.4.5	Pleasure, aesthetics, and usability	22
3.6	Conclusions	26
4	Context-generic psychometric tests of emotion and affect	28
4.1	Genva Appraisal Wheel	28
4.2	The Profile of the Non-Verbal Sensitivity (PONS)	30
4.3	Diagnostic Analysis of Non Verbal Accuracy (DANVA)	31
4.4	Positive and Negative Affect Scale (PANAS)	33
4.5	The Self Assessment Manikin (SAM)	35
4.6	UWIST Mood Adjective Checklist (UMACL)	36
4.7	Differential Emotions Scale	36
5	Psychometric tests of usability	38
5.1	ASQ (After Scenario Questionnaire)	38
5.2	IsoMetrics	38
5.3	PSSUQ (Post Study System Usability Questionnaire)	39
5.4	QUIS (Questionnaire for User Interface Satisfaction)	40
5.5	SUMI (Software Usability Measurement Inventory)	42
5.6	WAMMI	43
5.7	Conclusions	44
	References	46

---

# 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 (in combination with Deliverable D9g) 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 examine the ‘state of the art’ with respect to the application of psychometric measures to emotion-related systems and to propose directions for future research.

- To catalogue and evaluate a range of psychometric measures suitable for the assessment of aspects of usability of emotion-related computing systems. This will include descriptions of the psychometric properties of these instruments, where this information is available.

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. This report is intended to provide researchers, from a range of academic and technical backgrounds with information that will assist in applying psychometric techniques in studies of emotion-related computing systems.

### 1.3 Main elements of the deliverable

By way of ‘setting the scene’ in Section 2 of this report we provide a brief overview of some key topics and issues that bear directly on the application of psychometric assessments in the context of the usability of emotion-related computing systems. This includes defining emotion-related computing systems; considering different models of emotion and affective states; describing the nature of usability testing; and reviewing the fundamental principles of psychometric testing.

Following this, in Sections 3, 4, and 5 of the report, we present material on three distinct areas of application of psychometrics in this context. In Section 3, we review empirical work concerned with the development of new psychometric measures that are specifically focused on users’ affective responses to computing systems. Some of these studies have used data reduction techniques to identify component elements of these responses and as a basis for model development. In Section 4 we review a range of context-generic psychometric assessments of affect that are well-founded in the psychological literature, and that have been or could be applied in the context of emotion-related computing systems. We describe the purpose of each measure and its psychometric properties. Finally, in Section 5, we describe a range of psychometric measures of ‘global’ usability. These are well recognised measures that are widely used in studies of usability, but that may not have been used in the context of emotion-related computing. Again, we describe the nature of these assessments and their psychometric properties. Finally, in Section 6, we draw conclusions and consider prospects for future development of psychometric testing in this specific context.

**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. Setting the Scene

In this section we present some brief explanatory material relating to the nature of emotion-related computing systems, the structure of emotions, the nature of usability testing, and the principles of psychometric assessment. The first three of these topics are covered in more detail in HUMAINE project deliverable D9g.

### 2.1 Emotion-related computing systems

The HUMAINE project is concerned with computing systems that can “register, model and/or influence human emotional and emotion-related states and processes” (HUMAINE project summary). Descriptive terms found in the literature for such systems include ‘affective computing’ systems (see Picard, 1997) and systems concerned with ‘User Experience’ (UX: Hassenzahl and Tractinsky, 2006). In Deliverable D9g, we argue that affective computing systems form a subset of systems designed for user experience. A key distinguishing feature of affective computing systems is their adaptability to the user, changing dynamically according to the user’s emotional state. Systems that do not adapt may still have been designed with the user’s affective state in mind. For example, an internet banking application could be designed on the basis of *a priori* assessments of interface characteristics (e.g., colour combinations) to elicit feelings of trust in the user (see e.g., Kim & Moon, 1998).

In this report we refer to ‘emotion-related systems’ to indicate membership of the general set. The deliverable is not restricted to consideration of a particular type of emotion-related system. Differences in system type are simply noted and evidence is presented relating to the application of psychometric tests in all contexts.

### 2.2 The characteristics of emotion

There are differing opinions on the characteristics of emotion. Again, these issues are discussed in more detail in project deliverable D9g, and this is not repeated in full here. However, when considering psychometric assessment two variations in description are particularly pertinent. The first is the familiar distinction between descriptions of emotion in terms of a small number of dimensions (e.g., Russell, 2003; Scherer, 2005) as opposed to a set

of basic types (e.g., Ekman, 1992). The specifics of the dimensions and also the basic types are still the subject of ongoing research and debate. However, it is particularly pertinent, in the context of this report, to note that dimensional models of emotion have tended to be based on psychometric assessments, whereas models of basic emotions have arisen from assessments of behavioural and physiological changes (Power, 2006).

The second variation in description concerns the level of *granularity* at which affective responses are considered. This can apply to both dimensional and typological models of emotion. In both cases it is possible to consider description as being somewhat hierarchical, in which each of a small number of emotional dimensions, or basic types, can be deconstructed to produce more detailed and less generic components (see e.g., Sloman, 2002).

### 2.3 The Nature of Psychometric Tests

On the basis that readers of this deliverable come from a wide variety of backgrounds, we provide here a brief introduction to principles of psychometric testing. This includes an explanation of the key test attributes of reliability and validity. The interested reader can find more information on these topics in a variety of good texts on psychometrics, including Kline (1999, 2000) and Nunnally & Bernstein (1994). Annett (2002) provides an evaluation of self-report rating scales with specific reference to application in the domain of ergonomics.

Psychometric principles can be applied to data of various types. However, in this report we focus exclusively on self-report data and the term ‘psychometrics’ is only used to refer to this. Self-report psychometric tests typically comprise a number of items (questions) to which the test taker must respond. Test items are selected so that aggregation of item scores provides a standardised measurement of one (in the case of a test with a single scale) or more (in the case of a test with multiple scales) psychological constructs (e.g., pleasure). A fundamental premise of psychometric testing is that individuals possess a ‘true score’ on any given psychological construct, and that with a perfect method of measurement it would be possible to identify this true score. Unfortunately, we must accept that available psychometric measures do not reach this standard and, depending on the quality of the test, may or may not give a good approximation of the true score. On the assumption that variance in error (the distance between each item measurement and the ‘true score’) is random, it follows that the average of many repeated measurements of a single construct will tend to produce a more accurate assessment of true score than the average of few assessments.

This is the basis for calculating test *reliability*, a statistic that provides information on the accuracy of a test. Correlations between item scores form a basis for judging whether all the items in a scale are assessing the same psychological construct. A strong correlation would indicate that this is so. An assessment of scale length contributes information on the likely magnitude of error variance with, as described above, longer scales suggesting less error (because scores from a greater number of items are aggregated to provide a total score). When these two components are combined a reliability coefficient is produced that is in the range 0 to 1. A coefficient of 1 indicates perfect reliability. Generally, if a test (scale) has a reliability coefficient of less than 0.7 this is a cause for concern. In such cases there is an undesirable amount of error in measurement.

*Validity* is a second important quality to consider when evaluating psychometric tests. Validity is concerned with whether a test is actually assessing the intended psychological construct. A test could be found to be reliable, but still not assessing its intended target. Marshalling evidence to support test validity is more complex than for reliability, with evidence coming in different forms:

- i) *Face validity*: Do the items in a test appear to measure the intended psychological construct? Face validity can be an important determinant of participants' willingness to complete psychometric tests. However, high face validity of items can also leave tests more susceptible to response manipulation by participants.
  
- ii) *Criterion-related validity*: Do scores on the test/scale correlate with measures of similar psychological constructs? In the context of emotion-related computing, a common approach is to gather psychometric and psychophysiological, e.g., Galvanic Skin Response (GSR), Heart Rate (HR) assessments in the one study to corroborate measurements of experimental manipulations. Often this is so that valence of emotion can be determined, given that some psychophysiological measures (e.g., GSR) provide indicators of arousal but not valence. However, this approach can also be considered to contribute to the process of validating psychometric assessment and contributes to determination of construct validity (see below).

- iii) *Content validity*: Do the items in the scale reflect all aspects of the psychological construct under investigation? Content validity is always difficult to establish. A number of studies (see Section 3) have used statistical data reduction methods to provide evidence as to the structure of users' affective responses to system designs. However, if the items selected for testing are not representative of all aspects of emotional experience, or the testing scenarios do not reflect an appropriate range, it can come as no surprise if statistical analysis of the structure of responses does not identify all possible emotional constructs (see Annett, 2002). Moreover, when using statistical data reduction techniques (e.g., factor analysis), an imbalance in the sampling of items, with respect to the constructs under investigation, is likely to reduce the validity of the resulting solution. There is no simple formula that can be applied to assess content validity (unlike reliability) and it is usually left to expert judgement. However, in the case of emotion, issues of content validity extend beyond the selection of items. It has been empirically demonstrated that emotional responses do not necessarily require conscious, effortful processing and will not necessarily be available to conscious introspection (Berridge & Winkielman, 2003). This presents difficulties for the psychometric approach in providing an all-encompassing assessment.
- iv) *Construct validity*: Do experimental manipulations lead to psychometric test results that are consistent with theory? When considering the application of psychometric testing in the context of emotion-related computing systems, the debate as to whether emotion should be characterised in terms of dimensions or types (see above) presents potential difficulties with regard to construct validity. The different theoretical positions can impact the selection or design of psychometric measures. However, much of the research in this area has ignored the 'bigger picture' and, instead, has focused on more detailed descriptions of affective states that are considered particularly pertinent to human-computer interaction (e.g., frustration, flow). How these affective constructs fit in the context of broader theoretical frameworks is not altogether clear. For example '*frustration*' might be a marker for an emotional dimension of obstructive-conductive (see e.g.,

Scherer, 2005) or a pre-emotional state that relates to the basic emotion of anger (Bessière, Newhagen, Robinson, & Shneiderman, 2006). Further research is required that will help to bridge the gap in descriptive terminology between applied and theoretical research. We return to this issue later.

In selecting methods of measurement for ergonomics studies, Annett (2002, p. 980) suggests that “the first question to ask is whether subjective experience of the user/operators is relevant to the purpose of the study”. In the case of studies of emotion-related computing systems it is difficult to imagine this not being the case. However, this is not to say that self-report measures are sufficient. Again, this is a point that we return to.

## 2.4 Usability testing

Perhaps the most recognised definition of usability is provided by ISO 9241. Here usability is defined as 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.” (page 2).

- *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.

A recent paper by Hornbaek (2006) demonstrates that there has been significant activity in this area since the publication of this ISO standard in 1998. In a review of papers published in six journals or collections of proceedings between 1999 and 2002 he identifies 180 articles that report the use of measures of usability. More recently a broader perspective has been suggested that incorporates extensive consideration of the motivation of the user (Dillon, 2001). The construct of ‘satisfaction’ provides a limited description of the affective response of the user to a given computer system (see Section 3).

When considering the process of developing and operating usable emotion-related computer systems, psychometric tests can be employed in several different ways. They can be used to assess users' emotional responses to computer interfaces in an experimental context as part of the development process. This information can provide a basis for the design of 'static' systems. For example, it might be used to determine whether users regard a particular ecommerce interface design as being trustworthy. Alternatively, psychometric assessment might be used to validate concurrently measured behavioural and/or psychophysiological changes that it is hoped will be sensitive to specified interface manipulations (e.g., the sensitivity of GSR to manipulations designed to induce frustration in the user (see Mahlke, Minge, & Thürling, 2006). These behavioural and psychophysiological measures have greater temporal resolution and are less intrusive on task performance, and are therefore more likely to be used as part of a dynamic affective computing system than psychometric assessments. However, although difficult, it is not unthinkable to integrate psychometric assessments in dynamic systems. Depending on the nature of the task it may be possible for users to make continuous or periodic responses to questions about affective experience (e.g., Cowie et al., 2000). Finally, psychometric tests are available to provide direct assessments of usability, including evaluation of the effectiveness, efficiency or performance with the system. In this report we describe some of these and consider how well they are suited to the evaluation of emotion-oriented systems (see Section 5). It is not necessarily the case that assessments that have proved successful in other settings will be equally successful in this domain. There are usability issues that are unique to emotion-oriented computing systems. For example, usability might relate to the efficacy with which the user's emotions are (dynamic) or have been (static) assessed by the computing system/design and the value that system users derive from that.

### 3. Computing-specific psychometric tests of emotion/affect

In this section of the report we review research concerned with the development of psychometric tests for the assessment of affective responses to computing systems and/or research efforts to develop models of affective response from psychometric (self-report) data. We begin by considering the construct of user satisfaction, before examining a range of other affective and attitudinal constructs. The role of aesthetics is discussed, and the use of statistical data reduction techniques as a means of identifying key constructs of user response is described and evaluated.

#### 3.1 Satisfaction: Affect or attitude?

As described in Section 2.3, user satisfaction is a key element of ‘traditional’ conceptualisations of usability assessments (ISO 9241) and is typically assessed using self-report (see Hornbæk, 2006). Satisfaction is held to reflect the user having a positive attitude towards a computer system and being free from discomfort (ISO 9241). This raises the issue of the distinction between attitudes and affect, and is relevant to a number of other computer-related constructs that are assessed using psychometric measures (some of these are discussed later in this section). As with emotions, attitudes are directed ‘at something’; they “... refer to people’s global evaluations of any object, such as oneself, other people, possessions, issues, abstract concepts, and so forth” (Petty, Fabrigar, & Wegener, 2003, p. 752). Current psychological models of attitudes hold that they are determined on the basis of three components, cognition, emotion, and behaviour (see e.g., Petty, Fabrigar, & Wegener, 2003). So, users’ attitude will be influenced by their affective response to a computer system, but affective response will not be an exclusive determinant. When considering the process of a computer user making self-report assessment of affect, as discussed in Section 2, this generally interferes with task performance. Therefore responses are usually solicited either periodically or at the end of an interactive session. This means that the users’ assessment of satisfaction, will be determined by an evaluation of: i) affect over the period; ii) the correspondence between actual and goal states, and, iii) behavioural tendencies. Following from this, perhaps we should not be surprised if the construct of satisfaction is correlated, at least to some degree, with other usability constructs, especially, effectiveness and efficiency (Sauro & Kindlund, 2005; although see Frøkjær, Mertzum, & Hornbæk, 2000). Self-report measures of satisfaction sometimes explicitly refer to effectiveness and efficiency of the system (e.g., “Are you satisfied with the ease of use of the system?”) (see Hassenzahl, 2004;

Hornbæk, 2006), and this is consistent, in this context, with user satisfaction being assessed as an attitude rather than an emotion. See deliverable D9g for further discussion of some of these issues.

### 3.2 Standardising the measurement of user satisfaction

Many different measures of satisfaction have been used in empirical studies. These tend to be rather basic (Hornbæk, 2006; Lindgaard & Dudek, 2003). In his review of 180 studies of computing system usability, Hornbæk (2006) identified only 12 studies that had used standard questionnaire. The measures used included the Questionnaire for User Interface Satisfaction (QUIS: see Section 5) and components from Davis' (1989) questionnaire measures relating to 'technology acceptance' that focuses on 'perceived ease of use' and 'perceived usefulness'.

Early efforts to standardise the assessment of user satisfaction with computing systems focused on occupational settings and were based on models of job satisfaction (e.g., Bailey & Pearson, 1983). Researchers were concerned with increasing the uptake of computer systems, and were particularly driven by the move to end-user computing. The emphasis was firmly on increasing productivity in the work environment. Questionnaire items addressed system characteristics that may or may not produce user satisfaction and, in this respect can be contrasted with more generic assessments of emotion that are more introspective in nature (see Section 4). For example, Bailey and Pearson (1983) developed an extensive measure of user satisfaction. Users responded to 39 factors using four semantic differentials for each. Items required users to appraise various aspects of system (including support staff) performance.

A linking thread of relevant and sustained research activity has been conducted in the context of information systems. An early interest in user satisfaction in this domain may have resulted from the existence of information systems that incorporated human assistance for the searcher, rather than end-user computing. Ives, Olson, & Baroudi (1983; Baroudi & Orlikowski, 1988) developed and tested a short form of the Bailey and Pearson (1983) scale that comprised 13 factors each with two semantic differential response scales. Subsequent analysis by Doll, Raghunathan, Lim, & Gupta (1995) generally supported the instrument, and identified user satisfaction as a second level factor with four contributing factors of: i) Electronic Data Processing (EDP) services; ii) EDP staff; iii) Information product; and, iv)

Knowledge or involvement. Unsurprisingly, given the nature of the items, factors are ‘outward looking’ and strongly related to the computing/task environment as opposed to introspective reflections of the users’ affective experiences (as mentioned above). Also within this information systems context, Doll and Torkzadeh (1988) drew on previous work, to develop a 12-item measure of end-user computing satisfaction (EUCS). As with the Bailey and Pearson (1983) measure, items generally referred to properties of the system (e.g., “Is the system accurate?”) and the achievement of user goals (e.g., “Does the system provide reports that seem to be just about exactly what you need?”) rather than the affective experiences of the user. This is illustrated by the nature of five factors identified from factor analysis: content, accuracy, format, ease of use, and timeliness (see Etezadi-Amoli & Farhoomand, 1991, for a critique, and Doll & Torkzadeh, 1991, for a reply). More recently, Wang, Tang, & Tang (2001) developed a 21 item measure for application in the context of customer information satisfaction with digital products and services. Factor analysis identifying seven factors: i) customer support; ii) security; iii) ease of use; iv) digital products/services; v) transaction and payment; vi) information content; and, vii) innovation.

### 3.3 Satisfaction as a psychological and design construct

An important issue to be considered with regard to the assessment of user satisfaction concerns the psychometric properties of associated assessment scales. An absence of discomfort is one of the determinants of user satisfaction, as defined in ISO 9241. This perhaps supports the view that satisfaction is a rather ‘neutral’ concept (Edwardson, 1998; Linaard & Dudek, 2003). Moreover, it may be that a stronger response is elicited from users when conditions for dissatisfaction are present than when those for satisfaction apply. This would be consistent with the view taken by Jordan (2000) that consumers have expectations that products will be easy to use, they do not regard this as unusual (see also Blythe & Wright, 2003). From a different perspective, Blythe and Hassenzahl (2003) suggest that satisfaction results from the confirmation of expectations, whereas pleasure results from unexpected deviations. See Deliverable D9g for further discussion.

Based on an analysis of the positioning of ‘satisfaction’ and a number of other emotion-related terms on a two-dimensional circumplex model of pleasure and arousal (see e.g., Russell, 2003), Edwardson (1998, p. 11) concluded that “It may indeed be far more useful to measure and understand customer happiness and customer anger as the primary exemplars of consumer experience rather than satisfaction. Satisfaction is a theoretical

construct used by researchers and business, yet it has a unique structure and meaning for customers themselves”. Edwardson found happiness and anger to be relatively neutral with regard to the arousal dimension but representing opposing extremes of the pleasure dimension, whereas satisfaction was a pleasant, low-arousal experience.

### 3.4 Extending the measurement of user affect

In recent years researchers and system developers have begun to take a broader view of users’ affective experiences when interacting with computer systems (see e.g., Blythe & Wright, 2003; Dillon, 2001; Hassenzahl & Tractinsky, 2006). This change coincides with an increase in the users’ opportunity to select between systems, and it seems probable that these two events are not unconnected. For example, the Internet often provides a range of sites offering similar facilities. If the user does not like one site they can quickly transfer to another. Computer systems are facing similar issues as consumer products that compete on a supermarket shelf. Also important are changes in the nature of computing systems. As Blythe & Wright (2003, p. XVI) suggest “... information and communication technology have moved out of the office and into the living room”. Users want a range of different facilities from computing systems. Edwardson (1998, p. 2) describes the potential impact on user experience rather well in relation to the hospitality, tourism, leisure, and entertainment industries: “We don’t ski to be satisfied, we want exhilaration”. Similar principles may apply in some computing contexts. Finally, as computer technologies advance there is the potential to provide users with richer and more immersive experiences that mirror many qualities of real life (and perhaps beyond!). This must also serve to extend the boundaries of user experience. It seems clear, in light of these developments, that assessments of user satisfaction will be insufficient. With this in mind, researchers have proposed and examined a variety of emotion-related constructs.

In contrast to studies that utilise psychophysiological assessment of user responses, many psychometric assessments of users’ responses to computing systems have focused on positive affective constructs (see Hassenzahl & Tractinsky, 2006). This is consistent with the studies reported in Hornbæk (2006). In his review of 180 studies of computing system usability, published in core HCI journals and proceedings between 1999 and 2002 (exact dates depend on source), Hornbæk (2006) identified 70 measures of specific users attitudes that had been assessed by self-report (see Table 3.1). Of these, 13 address explicitly negative emotional or physical states.

Measure	n	Explanation
Annoyance	7	Measures of annoyance, frustration, distraction, and irritation
Anxiety	3	Users' anxiety when using the interface
Complexity	3	Users' perceptions of interface complexity
Control	7	Users' sense of control and attitude towards the level of interactivity
Engagement	4	Users' experience of engagement, involvement and motivation
Flexibility	3	Users' perception of flexibility in the interface
Fun	14	Users' feeling of fun, entertainment, and enjoyment
Intuitive	3	Users' perception of the intuitiveness of the interface
Learnability	5	Users' attitude toward how easy it was to learn to use the interface
Liking	15	Users' liking of the interfaces
Physical discomfort	3	Users' experience of physical discomfort in using the interface
Want to use again	3	Users' attitude towards using the interface again

Table 3.1. Measures of specific attitudes towards the interface (from Hornbæk, 2006).

It is apparent that the theoretical grounding for some of these assessments has not been clearly established. They also vary with respect to whether they evaluate properties of the interface (e.g., complexity, flexibility) or experiences of the user (e.g., anxiety, engagement, fun). This is a point that we return to later when discussing factor analytic studies. In the following paragraphs we consider the psychometric assessment of some of these and other constructs.

### *3.4.1 Computer anxiety*

From the early 1980s, a concern with the uptake of computing systems and the possible exclusion of certain groups of the population drove the development of scales designed to assess attitudes to computers (e.g., Computer Attitude Measure: Kay, 1993; the Computer Attitude Scale: Lloyd and Gressard, 1984; the Computer Attitude Scale: Dambrot, Watkins-Malek, Siiling, Marshall, & Garver, 1985). The CAS comprises three scales: computer confidence, computer anxiety, and computer liking, all with good reliability coefficients (Lloyd & Gressard, 1984; Smith et al., 2007). The concept of computer anxiety was particularly prominent in the late 1980s. Some empirical studies of this topic (e.g., Good, 1982; Guynes, 1988; ) used well-known, standardised anxiety scales, such as the State-Trait Anxiety Inventory (STAI: Spielberger, Gorsuch, Lushner, Vagg, & Jacobs, 1993). However, a number of computer-specific scales were also developed (e.g., Charlton & Birkett, 1985; Heinssen, Glass, & Knight, 1987; Marcoulides, 1989; Raub, 1981). It may be that these scales have become rather dated, or are somewhat limited in the computing contexts to which they apply. Recently, Barbeite & Weiss (2004) concluded that there was not measurement equivalence between data they gathered from an internet sample and initial validation studies for several tests of computer anxiety and computer self-efficacy. They could not be certain whether this was due to differences in the nature of the sample or differences in the method of administration. However, they noted some problems with items from the original tests that seemed out of date given the current prevalence of computers. They went on to develop revised tests of computer anxiety and computer self-efficacy that would be appropriate for Internet users.

### 3.4.2 *Flow, Fun, and Playfulness*

The psychological construct of *Flow* is related to playfulness (see Reid, 2004; Webster, Trevino, & Ryan, 1993) and refers to a feeling of being enjoyably engrossed in a task, to the exclusion of distractions, with performance under comfortable control (see Csikszentmihalyi, 1975). Flow states are encouraged by tasks that are sufficiently, but not overly demanding. In the context of computer-based performance this is potentially a very desirable state, with users feeling a sense of engagement, interest, and an ability to cope. Consequently, several HCI researchers have explored the concept. Webster et al. (1993) factor analysed a 12-item psychometric assessment of flow that addressed four components: control, attention focus, curiosity, and intrinsic interest. Results indicated that an 11-item three-factor solution (combining curiosity and intrinsic interest) better fitted the data, although for most analyses the full 11-item scale was used. Consistent with hypotheses, flow

was found to be correlated with experimentation, perceived software flexibility, quality of interaction, and use of the software applications (spreadsheet and electronic mail).

Huang (2003) used the 12-item version (although one item was omitted following analysis) of the Webster et al. (1993) questionnaire to examine a three-stage model of user response that incorporated effects of complexity, novelty, and interactivity of website designs as antecedents of the four flow components that resulted in *utilitarian* or *hedonic* user assessments (see Section 3.4.6). Flow contributed to both hedonic and utilitarian assessments, with patterns emerging from website/flow component/web assessment associations (paths), including interactivity promoting control, curiosity, and interest, but not attention, which was negatively associated with website complexity.

The concept of fun has also attracted a good deal of research interest (see e.g., Blythe, Overbeeke, Monk, & Wright, 2003). Based on a theoretical review of the concepts, Blythe & Hassenzahl (2003) distinguish between *fun* and *pleasure*. They suggest that fun cannot be serious. This clearly has implications for the types of computer-based tasks that would be congruent with each of these specific affective responses (cf. Edwardson, 1998). Fun is associated with being distracted from ‘everyday’ concerns, whereas pleasure is associated with being absorbed in the task. In this regard parallels can be drawn between the constructs of pleasure and flow.

### 3.4.3 Trust and loyalty

Assessment and consideration of users’ feelings of trust and loyalty have become increasingly important with the rise in use of ecommerce and informational websites (e.g., health). Several studies have examined design factors that promote trust (e.g., Basso, Goldberg, Greenspan, & Weimer, 2001; Flavian, Guinaliu, & Gurra, 2006; Kim & Benbasat, 2006; Sillence, Briggs, Peter, & Fishwick, 2006; van der Heijden, Verhagen, & Creemers, 2003) and a number of models of trust have been developed (e.g., Corritore, Kracher, & Wiedenbeck, 2003; Pavlou & Fygenson, 2006; Sillence et al., 2006). Hassell (2005) makes a useful contribution by considering the relationship between trust and affect. He distinguishes between trust and trustworthiness, suggesting that most studies have focused on the latter. He goes on to propose an interesting, if somewhat underspecified, model that incorporates ‘perceived ease of use’ and ‘attitude toward using’ as components (see above). Consistent with results reported by Edwardson (1998) the importance of contextual factors is

emphasised. For example, Edwardson (1998) found the experience of ‘trust’ was a key feature of experience with banks but not for retail clothing or hospitality industries. Also broadly supportive of this model are findings that usability is an important determinant of trust (Corritore et al., 2003; Flavian et al., 2006).

An interesting series of studies that employed a psychometric approach to examine website (banking) design factors that were associated with users’ perceptions of trust was conducted by Kim and Moon (1998). Initially they identified 318 ‘emotional’ terms (some of these might more properly be considered cognitive associations, e.g., ‘*countrified*’, ‘*in vogue*’). A sample of participants used these as a basis for assessing cyber-banking interfaces (home pages only). A cluster analysis of this data identified 10 clusters of terms and representative terms were selected for each (total terms=40). In a second study, design elements that users found memorable or prominent were identified using a memory test and a sorting approach. In a third study, participants rated 12 interfaces (single screens) on each of the 40 adjectives. Factor analysis was used to identify common constructs. Seven factors were identified: attractiveness, symmetry, sophistication, trustworthiness, awkwardness, elegance, and simplicity. Kim and Moon (1998) then examined the association between the ‘trustworthiness’ factor and the design elements identified in Study 2. The design elements related to: i) the use of clipart; and, ii) colour. It would seem that design elements were assessed as being present or absent (dummy coded) and used as independent variables. Results suggested that interfaces using a three-dimensional presentation format with no animation were regarded as being most trustworthy. In a final study Kim and Moon (1998) developed two interface designs to engender high and low trust, respectively. Predicted differences on the four ‘trust’ scales (dependable, reliable, realistic, uniform) were apparent between the two interfaces.

Work in this area is still at an early stage. Researchers are examining associations between design parameters and assessment of trust/trustworthiness, as described. However, much remains to be done, including evaluation of the issue of context dependence, as described earlier.

#### 3.4.5 Frustration

User frustration has been considered in many studies of affective computing. However, assessment is often made using psychophysiological methods. In contrast, Bessière,

Newhagen, Robinson, and Shneiderman (2006) used a diary methodology to examine predictors. Participants completed a computer-based work session (content of their own choosing) while keeping a record of frustrating experiences. A number of demographic (e.g., age, computer experience) and dispositional variables (e.g., computer anxiety, computer attitudes) were recorded (exact details for the scales are not given) along with assessments of frustration levels associated with individual incidents and the overall session (again, exact details of these measures are not given). Results indicated that situational variables (e.g., importance of task, time to fix, time lost) were the strongest predictors of specific incident frustration, while dispositional variables were predictive of session frustration. This difference in immediate and longer term predictors may have implications for the assessment of usability. For example, it is possible that end of session assessments of user satisfaction will not provide an accurate summary of affective responses to specific incidents (see Section 3.1).

#### 3.4.6 *Pleasure, aesthetics, and usability*

A common distinction in the ‘user experience’ literature is between the hedonic/aesthetic and utilitarian qualities of computing system interfaces (see e.g., Hassenzahl et al., 2000; Huang, 2003, 2005; Voss et al., 2003; Westerman et al., *submitted*). This is consistent with the distinction made in the design world between form and function (see e.g., Lidwell, Holden, & Butler, 2003), and also with ‘traditional’ views of computing system usability in which assessments of the practical implications of using a piece of software (effectiveness and efficiency) are distinguished from assessments of user affective/attitudinal response (satisfaction). However, as with assessments of other usability components, the divergent validity of these constructs can be challenged.

Several studies have used a psychometric approach to isolate these constructs. For example, Huang (2005) factor analysed an initial pool of 42 semantic differential (see Osgood, 1969) items to identify hedonic and utilitarian aspects in users’ assessments of their most frequently visited website (n=229). On this basis two 11-item scales were proposed however, further analyses led to two 5-item scales, and subsequently, following data gathering with a further sample who assessed allocated ecommerce websites (n=583) to a hedonic scale comprising three items (fun-frustrating, enjoyable-unenjoyable, and interesting-boring) and a utilitarian scale comprising four items (safe-risky, ordered-chaotic, wise-foolish, and reliable-unreliable). The scales were relatively strongly correlated, indicating the

potential for a single scale solution, although some criterion-related assessments (e.g., association with a measure of flow) supported the value of the two scales.

A similar study was performed in relation to consumer products by Voss, Spangenberg, & Grohman (2003). An initial pool of 24 items, reflecting utilitarian and hedonic dimensions of consumer attitudes towards products/brands was tested (n=608). Voss et al. (2003) indicate that analyses supported the two dimensions and on the basis of subsequent data gathering and analysis identified two 5-item scales. However, from the reported item-factor loadings the two factors are not immediately distinct, and it would appear, as with Huang (2005) a relatively strong correlation exists between them. Huang (2005) suggests that this may arise because both constructs are strongly related to the requirements of users or, alternatively, because users' hedonic response is moulded by their utilitarian response.

An association between these constructs is also consistent with the results of Tractinsky, Katz, & Ikar (2000) who examined participants' responses to experimental manipulations of the design of an Automated Teller Machine (ATM). They report a correlation between pre-use assessments of aesthetics and usability, but also between post-use aesthetics and usability. They point to this as evidence of the importance of careful consideration of both constructs as part of the HCI design process. Lavie & Tractinsky (2004) extended this programme of work and developed a two factor psychometric assessment of aesthetics ('classical' and 'expressive'). A pool of 41 items was assembled and used as the basis for data gathering in Study 1, with participants (n=125) rating one of two websites (a news magazine or a usability organisation website) using 7-point Likert scales. A two-factor solution to the data was favoured, the item pool modified (35 items), and tested with a second sample (n=212) who visited an ecommerce website. Two factors were again identified, the first representing 'traditional' aesthetics, and the other representing more creative, 'expressive' aesthetics. A third study (n=145) examined the association between these factors and the constructs of usability, playfulness, pleasure, and service quality. Participants were required to assess a food and wine website. Results confirmed the two aesthetics factors. However, subsequent analysis of all items, including those representing the additional constructs did not provide a clear (simple) solution (some items loaded on more than one factor). In this context, participants found items relating to the 'playfulness' construct difficult to assess. In a final study (n=384) the two aesthetics scales were broadly replicated (although with some revision) and assessed in relation to usability, pleasure, and service quality scales

(these were initially identified independently of the aesthetics factors). Analyses indicated that the classical aesthetics factor was strongly correlated with all other constructs. However, correlations for expressive aesthetics were somewhat weaker (moderate correlations), and particularly so, for usability and service quality factors.

Hassenzahl (2004) also distinguishes between hedonic and utilitarian (*pragmatic quality*: PQ) aspects of design, and between two types of hedonic response. In Hassenzahl's model, '*Hedonic quality – identification*' (HQI) relates to the human need for self-expression in a social context. '*Hedonic quality – stimulation*' (HQS) relates to design novelty and the human need for change. Although not a complete match, some parallels can be drawn with the *classical* and *expressive aesthetics* constructs identified by Lavie and Tractinsky (2004). Hassenzahl (2004) describes a questionnaire measure of these three constructs (AttracDiff 2). However, it would seem that the development of this is reported in German, and so it is not discussed here. Hassenzahl (2004) presents a study in which this questionnaire was used as a basis for data gathering on different designs of MP3 player skins that varied in beauty/ugliness. '*Beauty*' and '*Goodness*' (similar to 'satisfaction') were also assessed. Using partial correlation to control for the effects of other attributes, HQI was the only questionnaire scale associated with *beauty*, whereas both HQI and PQ were associated with *goodness*. This does not support the link between aesthetics and usability as hypothesised by Tractinsky et al. (2000). Hassenzahl (2004) speculates that this may be due to Tractinsky et al.'s participants expressing their satisfaction with the interface when being asked about usability (see arguments above), or the absence of full factorial manipulation of aesthetics and usability in Tractinsky et al.'s studies. In a second study, Hassenzahl required participants to interact with the MP3 skins. The pattern of results was not clear as would be hoped, but HQI tended to be the scale most closely associated with *beauty*, and PQ was the scale most strongly associated with *goodness* following interaction. Hassenzahl suggests that goodness (satisfaction) is associated with both hedonic and utilitarian constructs, but that beauty is only associated with hedonic constructs.

Park, Choi, & Kim (2004) used a much larger pool of items (278 items) as the basis for analysis of users' responses to web page aesthetics and arrived at a more complex solution. Their goal was to examine the correspondence between designers' and users' impressions of web pages. Participants (n=418) rated four web pages on all adjectives. Cluster analysis identified 13 aesthetic components (bright, tense, strong, static, deluxe, popular, adorable, colourful, simple, classical, futuristic, mystic, and hopeful). A group of Web

designers then selected 30 adjectives, such that two or three items represented each cluster. Following this, Web designers were set the task of developing web pages that elicited one of the 13 aesthetic dimensions. During the design/development process verbal and action protocols were taken and these were analysed to identify key visual elements. In total 256 visual elements were identified and then allocated to the different dimensions (e.g., certain visual elements were held to elicit a 'bright' aesthetic impression). Finally, a sample of participants (n=515) rated the 52 web pages on the remaining 30 adjectives. Park et al. report that confirmatory factor analysis was successful in identifying the expected 13 dimensions and, following regression analyses, that there was good correspondence between the purpose-designed web pages and participants' assessments. However, from detail provided in the report, it is not entirely clear how strongly these conclusions are supported.

The Kansei Engineering approach is used mainly in the context of product design, but bears several similarities with the methods described here (see Nagamachi, 2002; Schütte, Eklund, Axelsson, & Nagamachi, 2004). The term 'Kansei' refers to the feelings experienced by the user in relation to a product, images they have in their mind of that product, and the impression they form of that product (Schütte et al., 2004). Kansei Engineering describes a set of methods that facilitate the application of this information to product design. Part of the typical Kansei process involves the generation of Kansei words that describe users' perceptions of, and affective responses to the domain (i.e., the range of possible design solutions). These are represented in the form of semantic differentials (see Osgood, 1969). Kansei Engineering is concerned with the process of mapping these Kansei words with different design alternatives. On this basis design solutions can be identified that produce desired Kansei (user sensations/feelings). These methods have been used successfully on a number of commercial products, such as the Mazda MX5 (Nagamachi, 2002). However, they do have drawbacks when considering the potential for theory generation. Although some Kansei words may be selected from those used in previous projects, typically a number are generated specifically for each product under investigation. This makes comparability of results difficult. It also creates problems with regard to the integration of results with existing theories of human affective responses to designs.

### 3.5 Conclusions

Many psychometric assessments of user satisfaction are reported in the literature. Hornbæk (2006) expresses concern over the number of different measures used and the lack of standardisation. There is a risk that researchers will use the same terms to refer to different psychological constructs and, on other occasions, use different terms to refer to the same construct (cf. Lindgaard & Dudek, 2003). In the context of personality assessment, Block (1995) refers these situations as the ‘jingle’ and ‘jangle’ fallacies. As a consequence comparability of results is difficult and progression of research is impeded. Moreover, there is a danger that, as a new, broader perspective on user experience and affective response becomes the norm for studies of human-computer interaction (e.g., Dillon, 2001; Hassenzahl & Tractinsky, 2006), these difficulties will be exacerbated. The affective constructs at the top of the new emotion-related computing research agenda (‘frustration’, ‘engagement’, ‘trust’) are not the same as those receiving most attention in the more general literature on emotion and affect. However, generic models of ‘approach’ and avoidance’ (e.g., Carver, 2006) may provide a useful and easy ‘bridge’ between important elements of the two literatures, with only minor realignment required.

Certainly more needs to be done to allow criterion-related and thereby construct validity to be demonstrated in the context of studies of emotion-related computing. The inclusion of standardised questionnaires, in addition to more specialised psychometric measures, is not necessarily a practical option. Administration times can become problematic, with participants becoming bored with the number of responses being asked of them. A sensible way forward would seem to be the identification and inclusion of a number of ‘marker’ items. These would be items from standardised tests or from other studies in the emotion-related computing literature that could be included as ‘reference points’ against which to judge new measures. Obviously these would not provide the reliability of full scales (see arguments in Section 2), but this would seem a reasonable compromise.

The other important way forward involves the use of behavioural and psychophysiological measures to support self-report (see e.g., Mahlke et al., 2006). Each type of measure has its advantages and disadvantages. If we regard these as being random between methods, the use of multiple methods and the aggregation of results can be expected to provide more reliable answers in much the same way that longer scales are more reliable than shorter scales. A major shortcoming of self-report measures, in this context, concerns their insensitivity to unconscious processing. It has been demonstrated that emotional responses of which the individual is unaware can influence behaviour (Berridge & Winkielman, 2003). A

better understanding of the circumstances in which this may apply and conditions of dissociation between cognitive and emotional determinants of behaviour (cf. Annett, 2002; O'Donnell & Eggemeier, 1986) would be valuable. Again, multiple assessment methods, including assessments of user performance/behaviour, seem to provide a way forward.

The use of data reduction techniques, for the purposes described in this section, can also present problems. We argue (Westerman et al., *submitted for review*) that, in some cases, insufficient attention is paid to the stages of processing and the types of construct under investigation (see Bloch, 1995; Doll and Torkzadeh, 1991). As a consequence, some factors that have been identified in published studies include items that relate to constructs the overall values of which may correlate but that are nevertheless conceptually distinct. For example, in a particular context the items 'bright', 'colourful', 'traditional', 'pleasing' might correlate and load on the same factor. However, in our view, they are conceptually distinct and should not be included in the same analysis (cf. Tractinsky et al., 2000).

## 4. Context-generic psychometric tests of emotion/affect

In this section of the report we review some of the key psychometric assessments of affect. This list is not exhaustive, by any means, but is intended to provide some points of reference that can be used when planning or evaluating studies in the area of emotion-related computing systems. With regard to selection criteria - Tests that are relevant to clinical samples only (e.g., relating to schizophrenia) have been omitted.

### 4.1 Geneva Appraisal Wheel

The Geneva Appraisal Wheel (GAW) has been recently developed by the Geneva Emotion Research Group and is reported in some detail in Scherer (2005). The purpose of the GEW is to measure the emotional reaction of participants to a variety of stimuli. These could be objects, situations, events or some kind, or interaction with a computer – such as using a new piece of software or visiting a new website. As such it represents a recent and useful non-invasive tool for the measurement of affective responses and may be well suited to applications within affective computing.

The emotion wheel consists of 20 ‘emotion families’ arranged in a circle around a central hub. The central hub consists of two further options that participants might select – that they felt no emotion whatsoever or that they felt some other emotion that is not listed on the wheel. From this central hub there are five circles, increasing in size moving out to the rim of the wheel for each of the emotion families. These are designed to represent the intensity of the emotion, with those circles further away from the centre representing a more intense emotion than those near the centre. The emotion families consist of two related affective terms such as ‘Happiness / Joy’ and ‘Disappointment / Regret’.

The emotion families are arranged around the wheel along two further appraisal dimensions. Running vertically through the wheel is the concept of Control / Power Appraisal from high at the top of the wheel to low at the bottom of the wheel. Running across the wheel are the appraisal dimensions of Pleasantness-Unpleasantness and Obstructive-Conducive.

On being presented with the emotion wheel participants are told that the two words that represent each family of emotions are merely indicative and can stand for a range of similar emotions. Thus as the instructions note the Irritation / Anger family can also covers

emotions such as rage, vexation, annoyance, indignation, fury, exasperation or being cross or mad. Participants are told that they must select the emotion family that is most similar to the emotion that they experienced when the 'event' happened. Following this they must indicate the intensity of the experience must selecting one of the increasingly large circles moving from the central hub to the emotion family, with the larger circles nearer the hub being the most intense. Participants may, if they wish, select two emotion families from the wheel and these emotions can be recorded as having been experienced at different intensities.

The Geneva Emotion Wheel is a relatively recent development in the measurement of emotion and it is apparent that the wheel itself is still under development. As such there is little currently available in terms of reliability and validity data. However, Banziger, Tran and Scherer (2005) reported some of the work that has begun to assess the validity of the tool. However, this reports information on an earlier version of the wheel that has 16 emotion families rather than 20. Using this version of the wheel they reported some difficulties around the task in terms of problems of differentiation, particularly with negative emotions. However, they found that the 64 adjectives used were generally assigned to the expected emotion family by their participants, thereby suggesting some degree of limited validity.

In a third task used by Banziger et al. (2005) participants were required to rate the 80 emotional categories / adjectives (16 categories or families and 64 adjectives – 4 for each family representing increasing intensity) along the dimensions of intensity, valence and control. This was done by using either visual analogue scales or by increasing or decreasing the size of a computer presented red circle to indicate the intensity of the experience whereas they had to click on a circle defined by valence (horizontal dimension) or control (vertical dimension) for any of the given emotional labels. It was found that there was only marginal support for the predicted trends here in terms of the progressive increase in intensity for the 4 adjectives in each category.

Consequently the GEW represents a potentially useful tool for the measurement of emotional responses and yet there is clearly scope for the tool to be tested and developed further at this stage. It seems that there needs to be far more testing for the validity of the tool before it can be used in a more widespread manner than is currently the case.

The Geneva emotion wheel is freely available for use for non-commercial research purposes and can be downloaded from <http://www.unige.ch/fapse/emotion> as can a copy of Scherer (2005).

#### 4.2 The Profile of Nonverbal Sensitivity (PONS)

The profile of nonverbal sensitivity (PONS) was developed by Rosenthal, Hall, DiMatteo, Rogers and Archer (1979). The scale was designed to increase understanding about the way that humans decipher any nonverbal information transmitted by another, whether that be by facial expression, body posture or the tone of voice in which a verbal message is delivered.

The PONS consists of a 45 minute film in black and white of 220 auditory and visual segments (some segments are audio only, others visual only and others still consist of both). These segments are a randomized presentation of 20 scenes all of them portrayed by a young woman. Each one of the scenes is represented within one of 11 'channels'. In terms of visual information there are three types of cue shown to participants, face cues, body cues and figure cues (consisting of both face and body). The auditory information consisted of a randomized spliced voice (RS) or an electronically content-filtered voice (CF). Combining the auditory and visual cues gives 6 audio visual channels. In addition to these there are 5 'pure' channels consisting of either visual information *or* auditory information (face, body, figure, RS and CF). Thus with the 20 scenes represented in each channel there are a total of 220 clips.

Participants must select one from a forced choice of two descriptions of what is being portrayed in the scene presented to them for example, they are shown one of the scenes and must decide whether the protagonist is: a) expressing strong dislike or b) ordering food in a restaurant. Participants were required to complete this for each of the 220 scenes portrayed in the film. It is important to note here that the voice filtering was designed to remove the content of the message but leave the tone of voice intact.

Rosenthal et al. (1979) report in detail issues regarding the reliability and validity of their measure, including reliability tests for each of the 11 channels. Internal consistency yielded a range of coefficients for 10 of the 11 channels of between .57 and .94. However, the randomized splice voice alone had a reliability coefficient of just .06. The reliability coefficient for the total test was .86. Thus the PONS shows a good level of internal

consistency. Test-retest reliability was also reported, with coefficients ranging between .18 and .52. Overall though the whole test has test-retest reliability coefficient of .69, it can be assumed then that the PONS has good stability reliability.

Rosenthal et al. (1979) also give details of the validity of the PONS test. Their treatment of this issue is rigorous and reported extensively throughout their book. They test the convergent and discriminant validity of the PONS using a plethora of other tests. For example they correlate the PONS with cognitive measures such as IQ, field independence (the ability to visually disembed hidden figures), other tests of nonverbal decoding (such as the CARAT, Buck, 1976; SIT, Archer and Akert, 1980) as well as a number of psychosocial correlates such as various personality tests (California psychological inventory, Gough, 1975; Marlowe-Crowne Social Desirability Scale (Crowne and Marlowe, 1964) and measures of interpersonal success.

While there is no space to consider these correlations in any detail here, it is worthy of note that Rosenthal et al. consider that the coefficients resulting from the testing described above fall close to the upper limit of what can be expected (somewhere around .30, Cohen, 1969) and the PONS also suitably low correlations where expected (such as IQ). They conclude then that the PONS demonstrates good criterion and discriminant validity.

#### 4.3 Diagnostic Analysis of Nonverbal Accuracy (DANVA)

Development of the Diagnostic Analysis of Nonverbal Accuracy (DANVA) was reported by Nowicki and Duke (1994). The DANVA was designed to measure individual differences in the ability to send and receive nonverbal social information in children aged between 6 and 10.

The DANVA was designed to measure four types of receiving ability through facial expressions, posture, gesture and tone of voice and three types of sending ability through facial expression, gesture and tone of voice. Thus the DANVA consists of 7 subscales.

The receptive facial expression test consists of 40 slides of faces (20 adult faces and 20 children's faces) showing 4 of each of the 4 emotions – happy, sad, angry and fearful plus 8 neutral face slides (4 children and 4 adult). The participant is simply required to classify the face into one of the four emotions or as 'other'.

The receptive posture test consists of 12 slides of a body posture (without the face shown) and the participant must classify them as happy, sad, angry or fearful. The receptive gesture test is presented in exactly the same way as the posture test with participants required to classify 12 slides as falling into one of the four emotions listed earlier.

The receptive voice test presents the participant with the same neutral sentence spoken to reflect each of the four emotions – happy, sad, angry or fearful. The test consists of 4 trials of each emotion making 16 in all. The role of the participant is to classify the tone of voice into one of the four emotions.

For the expression of emotion through facial features subtest participants are required to be filmed as they have a situation described to them and they must make a suitable facial expression that corresponds to the situation that has been described. There are two situations for each of the four emotions (happy, sad, angry and fearful) making a total of 8 descriptions in all. The facial expressions are then described as either being accurate or not for the situation by trained raters.

In the expressive gestures test the participants were filmed while they tried to describe by only using their hands. However, rather than asking for the four basic emotions listed above this time the participants were required to express some idea typically used to facilitate a social interaction, such as ‘inviting someone to play’ or ‘asking someone to stop.’ Again these films were rated by trained raters.

Finally, in the expressive language test the participants were required to read the neutral sentence ‘I am going to get my bike now and go for a ride.’ Before reading the sentence the participants had a situation described to them that was designed to elicit a certain emotional response from the four emotions (happy, sad, angry and fearful). Again these were rated by trained raters afterwards.

The DANVA shows good internal consistency with alpha coefficients for the receptive tests ranging from .77 to .88 and for the expressive tests they ranged from .68 to .82. It is clear then that the DANVA has good internal consistency. Nowicki and Duke also report test-retest reliability and these ranged from .70 (expressive language) to .86 (receptive gestures).

In terms of consistency Nowicki and Duke reported an impressive array of correlations with the DANVA. Suffice to say here that they reported that the ability to send and receive nonverbal information increased with age as was predicted. Nowicki and Duke also reported that the DANVA correlates significantly with a range of measures of personal and social adjustment and academic achievement. The DANVA does not correlate with measures of IQ but this is not necessarily surprising and is similar to the findings associated with the PONS outlined above.

#### 4.4 Positive and Negative Affect Scale (PANAS)

Watson, Clark and Tellegen (1998) reported on the development and validation of the Positive and Negative Affect Scale or PANAS. The test itself is a fairly simple and brief measure of affective state. The PANAS consists of 10 positive affect items (*attentive, interested, alert, excited, enthusiastic, inspired, proud, determined, strong and active*) and 10 negative affect items (*distressed, upset, hostile, irritable, scared, afraid, ashamed, guilty, nervous and jittery*). The scale can be used to assess affective states over 7 different time points ranging from 'at the moment' to 'during the past year' or 'generally'. Participants must rate each of the affective terms listed above in terms of the extent to which they 'feel this way right now' or 'felt this way during the past year' etc. using a five point scale ranging from 1 (very slightly or not at all) to 5 (extremely). The purpose of the PANAS was to provide reliable and valid positive affect and negative affect scales that had the added advantage of being both easy to administer and brief.

Watson, Clark and Tellegen (1998) report reliability scores for the PANAS using relatively large samples for each of the different time frames. For the 'moment' measure  $n = 660$ , 'today'  $n = 657$ , 'past few days'  $n = 1002$ , 'past few weeks'  $n = 586$ , 'year'  $n = 649$  and 'in general'  $n = 663$  (Watson et al. do not report data on the 'during the past week' time point). Also included are 101 participants who completed ratings on all seven time points on two occasions in order to provide test-retest reliability.

Internal consistency reliabilities of the PANAS are good with alpha reliabilities for the positive scale ranging from .86 to .90 and for the negative scale from .84 to .87 for the various time point measures. This suggests that there was little variation in internal consistency across the different time point measures. Watson et al. (1998) additionally report that there were low

correlations between the two scales with correlations ranging between  $-.12$  and  $-.23$  and as such are suitably low, suggesting that the two scales are relatively independent.

Test-retest reliabilities (computed using two administrations 8 weeks apart for every time frame, using a sample of 101 undergraduates) revealed higher correlations for the longer time frames, unsurprisingly ( $.68$  (PA) and  $.71$  (NA) for the ‘general’ time frame). However, even for the ‘moment’ and ‘day’ time frames the test-retest reliabilities were  $.54$  (PA) and  $.45$  (NA) and  $.47$  (PA) and  $.39$  (NA), suggesting that the PANAS also has good test-retest reliability.

In order to check the external validity of the PANAS Watson et al. reported correlations with a number of measures of related constructs (depression and anxiety). They report the correlations between the PANAS and 3 measures of associated constructs. Firstly, the Hopkins Symptom Checklist (HSCL; Derogatis, Lipman, Rickels, Uhlenhuth & Covi, 1974), the Beck Depression Inventory (BDI; Beck et al. 1961) and the State-Trait Anxiety Inventory State Anxiety Scale (A-State; Spielberger, Gorsuch & Lushene, Vagg & Jacobs, 1983).

The HSCL is a measure of general distress and dysfunction and correlates well with the negative affective scale of the PANAS -  $.74$  (past few days’ measure of PANAS) and  $.65$  (today). It shows slight negative correlations with the positive affective scale  $-.19$  (past few days) and  $-.29$  (today).

The BDI correlates with the PANAS NA scale ( $.56$  and  $.58$  past few days and today, respectively) as well as correlating negatively with the PANAS PA scale ( $-.35$  and  $-.36$  past days and today, respectively).

Finally, the STAI State Anxiety Scale also correlates with the NA scale of the PANAS – with a correlation of  $.51$  and also correlates negatively with the PA scale  $-.35$  (Watson et al., 1988 only report the correlations for the ‘past few days’ measure).

#### 4.5 The Self Assessment Manikin (SAM)

The Self Assessment Manikin (SAM) was developed by Lang (1980) and Hodes, Cook & Lang (1985) to address a number of issues with other measures of responses to

affective stimuli. The SAM is a picture-based assessment of affect designed to measure the valence, arousal and dominance of an affective response. By being picture based and requiring only 3 measures per stimulus presented it allows the measure to be used in situations where other measures are inappropriate (children, non-English speakers). This may make the tool valuable in settings with affective computing where internationalism and assessment involving various age groups may be important.

The pleasure dimension is represented by five images ranging from a smiling face to a frowning face. For the arousal dimension the figure ranges from wide-eyed and alert to sleepy. Finally the dominance dimension is represented by the changing size of SAM. Participants can place a cross over one of the figures or in between two figures, thus giving a 9-point rating scale. SAM also has the advantage of being presentable as either computer based or paper based.

Bradley and Lang (1994) tested the validity of SAM as a quick and efficient method of measuring affective responses. In order to do this they correlated ratings using SAM with the semantic differential scale (Mehrabian and Russell, 1974). Bradley and Lang report very high correlations between the SAM and the semantic differential scale for both the pleasure and arousal dimensions (.96 for pleasure and .95 for arousal). However, correlations were non-significant for the dominance dimension, although Bradley and Lang note that the differences in ratings of dominance may be due to the idea that SAM offers a better measure of control of the participant rather than the idea that there is dominance associated with the stimuli. For example, a picture of a snake may cause high ratings of dominance on the semantic differential (because the snake is dominant) whereas when using SAM participants rate their own dominance in the given situation (being faced with a snake) and thus the SAM yields a low dominance rating for such a stimulus. Ultimately, Bradley and Lang argued that SAM represents an accurate and efficient method for measuring affective responses to a range of stimuli.

#### 4.6 UWIST Mood Adjective Checklist (UMACL)

The UWIST Mood Adjective Checklist (UMACL) was proposed by Matthews, Jones and Chamberlain (1990) as a refinement to existing measures of mood. The UMACL consists of 24 adjectives which are rated along a 4 point scale as to whether they apply to the

participants' mood at moment of completion. The UMACL is designed to measure three dimensions – energetic arousal, tense arousal and hedonic tone. Matthews, Jones and Chamberlain (1990) reported on the validity of the UMACL as a measure of mood state. They reported low correlations between the UMACL scales suggesting discriminant validity. In addition to this they also reported that there were significant correlations between the arousal scales of the UMACL and various measures of arousal from psychophysiology which supports the notion of concurrent validity.

#### 4.7 Differential Emotions Scale

One of the most common tools used for the measurement of emotion is the Differential Emotions Scale (DES-IV) (Blumberg and Izard, 1986). Briefly the scale consists of 36 items on 12 scales and participants are required to state whether they have experience the emotion on a scale ranging from rarely to very often over the past week. The twelve scales that are used in the DES IV are *interest, joy, sadness, anger, disgust, contempt, fear, shyness, guilt, surprise, shame, and self-directed hostility*. Blumberg and Izard tested 280 children using the scale twice four months apart. They reported moderate correlations for the scales used, although there was considerable variation (correlations rated from .30 and .75).

This scale raises a number of important issues about the measurement of emotion. Firstly, it appears that what is being measured here is not emotion but rather mood (a longer lasting more stable affective event). Secondly there is the assumption that all groups of participants will approach the question in the same manner. Youngstrom and Green (2003) noted that there are often demographic differences when scales are used with different groups but that these are rarely considered by the researchers involved and there is an assumption that a validated tool will be valid for any group of participants.

#### 4.8 Basic Emotions Scale

Power (2006) reports the recent development of a measurement of emotion as a tool for the assessment of six different models of emotion. The scale consists of five subscales: *Anger, Sadness, Disgust, Fear and Happiness*. Three further emotion terms linked to each of these emotions were added to give a set of 20 emotion terms (these additional terms were taken from emotion terms provided by Johnson-Laird and Oatley (1989)). Participants were required to record how frequently they experienced each of these emotions using a 7 point

scale ranging from 'never' to 'very often'. Power reports the internal consistency for each of the five subscales and these are good with Cronbach's alpha scores ranging from .790 to .842. Thus it appears that the Basic Emotions Scale has good reliability. However, Power notes that there is a need for further validation studies of the scale and that these are yet to be completed. However, it appears that the scale does represent an interesting and relatively straightforward tool for the assessment of emotion.

## 5 Psychometric Tests of Usability

In this section of the report a selection of questionnaire-based assessments designed to gather data on overall system usability are described. We have selected those that we consider to be the most prevalent in the literature or on the web. Evidence bearing on the utility and psychometric properties of each is presented where it is available. These questionnaires were not designed for and have not been applied to emotion-related computing systems. An important question, therefore, is whether they will prove valuable in this context. The questionnaires are reported in alphabetic order of their acronyms.

### 5.1 ASQ (After Scenario Questionnaire)

This is a simple three-item questionnaire developed by Lewis (1991) to be used in scenario-based usability studies (i.e. the participants are asked to provide ratings based on their experiences completing tasks within particular scenarios, rather than report on their experiences of normal usage of the software). The items measure satisfaction with the ease of completing tasks, the amount of time taken to complete tasks and the support information available while completing tasks. Following psychometric evaluation, it was concluded that the items could usefully be summated to produce a single scale for measuring usability. Reliability was assessed using the internal consistency of the items with respect to each scenario (all coefficient alphas were reported in excess of 0.9). The scale was also shown to have good concurrent validity when compared to objective measures of scenario completion, such that those who successfully completed a scenario were more favourable in their ratings ( $r=0.4$ ).

### 5.2 IsoMetrics

The IsoMetrics inventory was originally developed by Gediga, Hamborg and Düntsch (1999) and was based on a comprehensive review of existing tools and literature, including the principles in ISO 9241 Parts 10, 12 and 14. A pool of 651 items was initially reduced to 151 and then subjected to further validation and reliability assessment. The result was a collection of 75-items covering the seven principles of dialogue design identified in ISO 9241-10. The seven principles covered, together with the number of items used in the current version are shown below:

- Suitability for the task (15)
- Self descriptiveness (12)
- Controllability (11)
- Conformity with user expectations (8)
- Error tolerance (15)
- Suitability for individualization (6)
- Suitability for learning (8)

The same 75 items are used in two versions of the scale, a long version which is recommended for evaluations which form part of the development process, and a short version which is recommended for use in evaluating systems following development. In the long version, respondents indicate their level of agreement with each item, e.g. “Too many different steps need to be performed to deal with a given task”, and their subjective impression of the importance of that item, using rating scales. Following these ratings, each item has a space for the respondent to give a concrete example where they can or cannot (different for different items) agree with the statement. In the short version, participants only provide the first agreement rating for each item.

In Gediga et al. (1999) the reliability of each sub-scale in the Beta-version is presented, which shows a range of Cronbach’s alphas from 0.69 to 0.86. This paper also reports a comprehensive account of how the two versions were validated with items being dropped to improve reliability. There is an IsoMetrics2 project underway that seeks to extend the validation of the inventory, develop a computerised form, widen the scope to include Multi-media and Hyper-media systems and develop a test manual. Some recent work by Hamborg, Vehse and Bludau (2004) in the area of Hospital Information Systems has shown that a comparison of online and paper-and-pencil presentations of the inventory do not significantly affect the ratings provided by participants.

### 5.3 PSSUQ (Post Study System Usability Questionnaire)

Like the ASQ, this is one of the IBM Corporation’s family of usability questionnaires. It has 19 items and is designed to assess “users’ perceived satisfaction with their computer systems.” (Lewis, 2002; p.464). The items are intended to address five system characteristics,

Quick completion of work, Ease of learning, High-quality documentation and online information, Functional adequacy and Rapid acquisition of productivity.

Following initial studies using the questionnaire, factor analysis techniques helped the authors to identify three main factors which now form the subscales of the questionnaire:

- i) System Usefulness
- ii) Information Quality
- iii) Interface Quality

Measures of reliability were most recently reported in Lewis (2002) for the overall scale as 0.96, and the three subscales as 0.96, 0.92, and 0.83 respectively. Criterion-related validity was reported in terms of a significant relationship between the PSSUQ and other measures of user satisfaction ( $r=0.80$ ). These figures were based on usability studies conducted with speech dictation systems over a period of five years and were compared with figures obtained 10 years previously. There were strong similarities between the two sets of data which Lewis suggests can be taken as evidence of the generalisability of the questionnaire.

There is an alternate form of the PSSUQ for use in field settings rather than a scenario-based usability evaluation, called the Computer System Usability Questionnaire (CSUQ). It has identical items to the PSSUQ although the wording has been made more appropriate for field settings. The measures of reliability were reported to be similar to the PSSUQ at 0.95 for the Overall scale, 0.93 for System Usefulness, 0.91 for Information Quality and 0.89 for Interface Quality.

#### 5.4 QUIS (Questionnaire for User Interface Satisfaction)

The development of this questionnaire was first reported in Chin, Diehl and Norman (1988). They describe how the first version consisted of 90 items, but was reduced in size through successive versions. The resulting version 5.0 comprised three sections; Section 1 dealing with the type of system, Section 2 concerning the user's past experience and Section 3 being the main section dealing with the user evaluation. Section 3 comprised 27 items covering 5 sub-scales:

- Overall reactions to the software

- Screen
- Terminology and System Information
- Learning
- System Capabilities

Each item is rated on a 9-point scale with anchored points labelled appropriately for each item (NB. Chin et al. (1988) describe the questionnaire as using a 1-9 scale but, in fact, the published paper presents a 10-point, 0-9 scale). Chin et al. (1988) go on to describe a study in which this short version of the QUIS (version 5.0) was given to participants who were asked to make comparisons between two pairs of software products (software that they liked vs. software that they disliked; a command line system vs. a menu-driven application). The reliability of the questionnaire was high (in common with previous versions) with a Cronbach's alpha of 0.94, and a factor analysis indicated a good (but not perfect) match between the sections and latent factors.

One of the suggested outcomes from this study was the creation of an online version of the questionnaire. Harper, Slaughter and Norman (1997) report on version 7.0, an expanded web-based version of the QUIS with four major sections:

- i) Demographics
- ii) Overall reaction to the system
- iii) Four interface factors (screen, terminology and system feedback, learning, system capabilities)
- iv) Optional sections (manuals and online help, online tutorials, multimedia, internet access, software installation)

The additional sections add a further 48 items to the questionnaire (each using the 1-9 scale, with negative adjectives corresponding to 1 and positive to 9). Harper et al. (1997) describe a validation of this version that revealed high reliability (Cronbach's alpha = 0.95) and good construct validity (correlations between the main items and a general satisfaction scale ranged between 0.49 and 0.61).

The advantages of a web-based version of this questionnaire include the facility to gather information from multiple users, in their own surroundings, with little data entry

overhead. Additionally, web-based delivery allows easy maintenance and the facility to extend or tailor its use.

### 5.5 SUMI (Software Usability Measurement Inventory)

This method for assessing usability was first introduced by Kirakowski and Corbett (1993). It comprises a 50-item inventory that draws on ISO 9241 principles and is recommended to be used with a minimum of ten users all of whom should have some experience of using the software to be evaluated. The inventory requires the user to either agree with, disagree with or state that they are undecided about each of the items. The psychometric development of the inventory is described in some detail in Kirakowski (1994) and this background paper describes how five subscales were identified (see below). In addition to these subscales, a Global scale of ‘general usability’ based on half of the items was also developed.

The overall reliability for these scales (Cronbach’s alpha) in the final version of the inventory is reported below alongside the descriptions of each sub-scale:

- Efficiency – the degree to which users feel that the software assists them in their work (alpha=0.81).
- Affect – general emotional reaction or ‘likeability’(alpha=0.85).
- Helpfulness – the degree to which users feel that the software is self-explanatory, including help systems and manuals (alpha=0.83).
- Control – a measurement of the feeling of control over the software that a user perceives versus the feeling that the software is in control (alpha=0.71).
- Learnability – how quickly the user felt that they were able to learn to use the system, or to acquire new skills (alpha=0.82).

The Global scale returned an alpha value of 0.92.

Kirakowski (1994) also reports a number of validity checks that demonstrate that SUMI evaluations consistently differentiate between so-called ‘friendly’ and ‘unfriendly’ versions of different types of software (e.g. word processors, databases).

SUMI is commercially available from the Human Factors Research Group in Ireland, but was originally developed as part of a European Commission funded project (Metrics for Usability Standards in Computing – MUSiC). One of the key features of the SUMI ‘package’ is that purchasers of the full version will receive computerised scoring facilities and a standardisation database. It is therefore possible to assess the usability of a piece of software by comparing its SUMI score with that of a competitor, or alternatively by comparing the score with an ‘average’ profile stored on the database for that type of software.

In addition to the questionnaires described above, there are a number of other methods for measuring usability that will not be dealt with in detail. This is because either there is no published data concerning reliability and validity for these scales (e.g., the System Usability Scale (SUS) by Brooke (unpublished) or the Usefulness, Satisfaction and Ease of Use (USE) questionnaire by Lund (2001), or because a scale is very similar to one by the same authors already described (e.g., the WAMMI Web Usability Questionnaire, which is very similar to SUMI).

## 5.6 WAMMI

“The WAMMI is a standardised measurement tool comprising 20 questions using 5-point response scales. It is claimed to be a valid and very reliable measure of perceived usability, including appeal of websites. The WAMMI is a measure of ‘global satisfaction’ and it is said to tap five dimensions of the user experience, namely (1) attractiveness (“the pages on this web site are very attractive”), (2) control (“I feel in control when I am using this web site”), (3) efficiency (“I feel efficient when I am using this web site”), (4) helpfulness (“this web site helps me find what I am looking for”), and (5) learnability (“learning to find my way around this web site is a problem”).” (Lingaard & Dudek, 2003, p. 434).

## 5.7 Conclusions

It is clear from the description of the questionnaire methods in this section of the report that there is a great deal of similarity between the methods and that perhaps the best way to discriminate between them will be based on the type of evaluation that is desired. For

instance, researchers will need to decide whether they wish to evaluate software in the normal working environment with little control over the tasks that the users will undertake or whether they wish to provide scenarios with more controlled tasks. Equally some of the questionnaire methods will lend themselves better to particular contexts or types of software.

One issue that is a constant concern for researchers is the length of time that the completion of the questionnaires will take since multiple evaluations may be required, e.g. to compare different systems or versions, or the length of the evaluation method may disrupt the normal use of the software thereby obfuscating the outcome of the evaluation. Some researchers have therefore questioned whether it would be possible to create very short evaluation methods based on a belief that the different factors in Usability may be highly correlated.

Frokjaer, Hertzum and Hornbaek (2000) take the three main aspects of usability as defined by the ISO, effectiveness, efficiency and satisfaction, and argue strongly that each factor should be taken account of in usability evaluations. They point out that, for instance, the execution of complex tasks seems to rely less on a need for efficiency and more on a need for effectiveness, suggesting that, in this context, it is necessary to take account of both factors. In fact, Frokjaer et al. (2000) go on to present data that suggests there are only weak correlations between the three factors and that concentrating on only subsets of these usability factors may lead to unreliable results. They point out that a high proportion of usability studies reported in the CHI Proceedings over recent years have not taken account of all aspects of usability and as such may be drawing conclusions about usability based on only some of the relevant information.

More recently Sauro and Kindlund (2005) report a method to standardize usability metrics into a single score. They argue that there is a strong need for usability evaluation methods to be easier and less complicated to use “to increase the meaningfulness and strategic influence of usability data” (page 401). They present data based from a two-year project in which four usability tests were administered to users of software from the financial and accounting industry. They demonstrate that it was possible to produce a summated usability metric (SUM) that summarized the variance in four scales that are often used in measuring usability. Although they do not claim that such a simple method should be used in all situations and certainly not to replace more in-depth methods during development of software,

they do claim that in summative evaluation a single continuous variable measuring usability may be useful.

## References

- Annett, J. (2002). Subjective rating scales: Science or art? *Ergonomics*, *45*, 966-987.
- Archer, D., & Akert, R. (1980). The encoding of meaning: A test of three theories of social interaction. *Sociological Inquiry*, *50*, 393-419.
- Baenziger, T., Tran, V. & Scherer, K.R. (2005). "The Emotion Wheel. A Tool for the Verbal Report of Emotional Reactions", Poster presented at the conference of the International Society of Research on Emotion, Bari, Italy.
- Bailey, J.E. & Pearson, S.W. (1983). Development of a tool for measuring and analysing computer user satisfaction. *Management Science*, *29*, 530-545.
- Barbeite, F.G. & Weiss, E.M. (2004). Computer self-efficacy and anxiety scales for an Internet sample: Testing measurement equivalence of existing measures and development of new scales. *Computers in Human Behavior*, *20*, 1-15.
- Baroudi, J.J. & Orlikowski, W.J. (1988). A short-form measure of user information satisfaction: A psychometric evaluation and notes on use. *Journal of Management Information Systems*, *4*, 44-59.
- Basso, A., Goldberg, D., Greenspan, S., & Weimer, D. (2001). First impressions: Emotional and cognitive factors underlying judgements of trust e-commerce. In *Proceedings of EC'01*, October 14-17, Tampa, Fl. New York: ACM Press. Pp. 137-143.
- Beck, A.T., Ward, C. H., Mendelson, M., Mock, J. & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry* *4*, 561-571.
- Berridge, K.C. & Winkielman, P. (2003). What is an unconscious emotion? (The case for unconscious "liking"). *Cognition and Emotion*, *17*, 181-211.
- Bessière, 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.
- Bloch, P.H. (1995). Seeking the ideal product form: Product design and consumer response. *Journal of Marketing*, *59*, 16-29.
- Block, J. (1995). A contrarian view of the five-factor approach to personality description. *Psychological Bulletin*, *117*, 187-2115.
- Blumberg, S.H. & Izard, C.E. (1986). Discriminating Patterns of Emotions in 10- and 11-Year-Old Children's Anxiety and Depression. *Journal of Personality and Social Psychology*, *51* (4), 852-857.

- Blythe, M. & Hassenzahl, M. (2003). The semantics of fun: Differentiating enjoyable experiences. In M.A. Blythe, K. Overbeeke, A.F. Monk, & P.C. Wright (Eds.), Funology: From Usability to Enjoyment. Dordrecht: Kluwer Academic Publishers.
- Blythe, M.A., Overbeeke, K., Monk, A.F., & Wright, P.C. (Eds.) (2003). Funology: From Usability to Enjoyment. Dordrecht: Kluwer Academic Publishers.
- Blythe, M. & Wright, P. (2003). From usability to enjoyment. In M.A. Blythe, K. Overbeeke, A.F. Monk, & P.C. Wright (Eds.), Funology: From Usability to Enjoyment. Dordrecht: Kluwer Academic Publishers.
- Bradley, M.M. & Lang, P.J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. Journal of Behavior Therapy and Experimental Psychiatry, 25 (1), 49-59.
- Brooke, J. (unpublished). SUS – A quick and dirty usability scale. Retrieved from <http://www.usabilitynet.org/trump/documents/Suschapt.doc>.
- Buck, R. (1976). A test of nonverbal receiving ability: Preliminary studies. Human Communication Research, 2, 162-171.
- Carver, C.S. (2006). Approach, avoidance, and the self-regulation of affect and action. Motivation and Emotion, 30, 105-110.
- Charlton, J.P. & Birkett, P.E. (1995). The development and validation of the computer apathy and anxiety scale. Journal of Educational Computing Research, 13, 41-59.
- Chin, J.P., Diehl, V.A. and Norman, K.L. (1988). Development of an Instrument Measuring User Satisfaction of the Human-Computer Interface. Proceedings of SIGCHI '88, 213-8. New York: ACM/SIGCHI.
- Cohen, J. (1969). Statistical power analysis for the behavioral sciences. London: Academic Press.
- Corritore, C.L., Kracher, B., & Wiedenbeck, S. (2003). On-line trust: Concepts, evolving themes, a model. International Journal of Human-Computer Studies, 58, 737-758.
- Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M. (2000). 'Feeltrace': An instrument for recording perceived emotion in real time. In Proceedings of the ISCA Workshop on Speech and Emotion, Belfast. Pp. 19-24.
- Crowne, D. P. & Marlowe, D. (1964). The Approval Motive. New York: John Wiley & Sons.
- Csikszentmihalyi, M. (1975). Beyond Boredom and Anxiety: The Experience of Work and Play in Games. San Francisco: Jossey Bass.

- Dambrot, F.H., Watkins-Malek, M.A., Silling, S.M., Marshall, R.S., & Garver, J. (1985). Correlates of sex differences in attitudes towards and involvement with computers. Journal of Vocational Behavior, *27*, 71-86.
- Davis, F.D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, *13*, 318-340.
- Derogatis, L.R., Lipman, R., Rickels, K., Uhlenhuth, E.H. & Covi, L. (1974). The Hopkins Symptom Checklist (HSCL): A self-report symptom inventory. Behavioural Science, *19*, 1-15.
- 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.
- Doll, W.J., Raghunathan, T.S., Lim, J-S, & Gupta, Y.P. (1995). A confirmatory factor analysis of the user information satisfaction instrument. Information Systems Research, *6*, 177-188.
- Doll, W.J. & Torkzadeh, G. (1988). The measurement of end-user computing satisfaction. MIS Quarterly, *12*, 259-274.
- Doll, W.J. & Torkzadeh, G. (1991). Issues and opinions – The measurement of end-user satisfaction. MIS Quarterly, *15*, 5-10.
- Edwardson, M. (1998). Measuring consumer emotions in service encounters: An exploratory analysis. Australasian Journal of Market Research, *6*, 34-48.
- Etezadi-Amoli, J. & Fahoomand, A.F. (1991). On end-user computing satisfaction. MIS Quarterly, *15*, 1-4.
- Flavian, C., Guinaliu, M., & Gurrea, R. (2006). The role played by perceived usability, satisfaction, and consumer trust on website loyalty. Information & Management, *43*, 1-14.
- Frokjaer, E., Hertzum, M. and Hornbaek, K. (2000). Measuring Usability: Are Effectiveness, Efficiency, and Satisfaction Really Correlated? Proceedings of ACM Conference on Human Factors in Computer Systems, *2(1)*, 345-352. New York: ACM Press.
- Gedgia, G., Hamborg, K.C., & Düntsch, 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.
- Good, M. (1982). An ease of use evaluation of an integrated document processing system. In Proceedings of the 1982 Conference on Human Factors in Computing Systems. New York: ACM Press.
- Gough, H. G. (1975). Manual for the California Psychological Inventory. Palo Alto, CA: Consulting Psychologists Press

- Guynes, J.L. (1988). Impact of system response time on state anxiety. Communications of the ACM, 3, 342-347.
- Hamborg, K-C., Vehse, B. and Bludau, H-B. (2004). Questionnaire Based Usability Evaluation of Hospital Information Systems. Electronic Journal of Information Systems Evaluation, 7(1), 21-30.
- Harper, B., Slaughter, L. and Norman, K. (1997) Questionnaire administration via the WWW: a validation & reliability study for a user satisfaction questionnaire. Paper presented at WebNet '97, Association for the Advancement of Computing in Education. Toronto, Canada.
- Hassell, L. (2005). Affect and trust. Lecture Notes in Computer Science, 3477, 131-145.
- Hassenzahl, M. (2004). The interplay of beauty, goodness, and usability of interactive products. Human-Computer Interaction, 19, 319-349.
- 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.
- Heinssen, R.K., Glass, C.R., Knight, L.A. (1987). Assessing computer anxiety: Development and validation of the Computer Anxiety Rating Scale. Computers in Human Behavior, 3, 49-59.
- Hodes, R., Cook III, E.W. & Lang, P.J. (1985). Individual differences in autonomic response: conditioned association or conditioned fear?. Psychophysiology, 22, 545–560.
- Hornbaek, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. International Journal of Man-Machine Studies, 64, 79-102.
- Huang, M-H. (2003). Designing website attributes to induce experiential encounters. Computers in Human Behavior, 19, 425-442.
- Huang, M-H. (2005). Web performance scale. Information & Management, 42, 841-852.
- ISO 9241-11. (1998). Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs). Part 11: Guidance on Usability. Geneva: International Organization for Standardization.
- Ives, B., Olson, M.H., & Baroudi, J.J. (1983). The measurement of user information satisfaction. Communications of the ACM, 26, 785-793.
- Johnson-Laird, P.N. & Oatley, K. (1989). The language of emotions: An analysis of a semantic field. Cognition & Emotion, 3 (2), 81-123.
- Jordan P.W. (2000). Designing Pleasurable Products. London: Taylor & Francis.

- Karat, J. (1997). User-centred software evaluation methodologies. In M. Helander, T.K. Landauer, & P. Prabhu (Eds.), Handbook of Human-Computer Interaction. 2<sup>nd</sup> Edition. Amsterdam: Elsevier Science.
- Kay, R.H. (1993). An exploration of theoretical and practical foundations for assessing attitudes towards computers: The Computer Attitude measure. Computers in Human Behavior, 9, 371-386.
- Kim, D. & Benbasat, I. (2006). The effects of trust-assuring arguments on consumer trust in Internet stores: Application of Tourmin's model of argumentation. Information Systems Research, 17, 286-300.
- Kim, J. & Moon, J.Y. (1998). Designing towards emotional usability in customer interfaces – trustworthiness of cyber-banking system interfaces. Interacting with Computers, 10, 1-29.
- Kirakowski (1994). The Use of Questionnaire Methods for Usability Assessment. Retrieved from <http://sumi.ucc.ie/sumipapp.html>.
- Kirakowski, J. & Corbett, M. (1993). SUMI: The software usability measurement inventory. British Journal of Educational Technology, 24, 210-212.
- Kline, P. (1999). The Handbook of Psychological Testing. Second Edition. London: Routledge.
- Kline, P. (2000). A Psychometrics Primer. London: Free Association.
- Lang, P.J. (1980). Behavioral treatment and bio-behavioral assessment: computer applications. In: J.B. Sidowski, J.H. Johnson and T.A. Williams (Eds). Technology in mental health care delivery systems. Norwood, NJ: Ablex.
- Lang, P.J. (1995). The Emotion Probe. American Psychologist, 50, 5, 372-385.
- Lavie, T. & Tractinsky, N. (2004). Assessing dimensions of perceived visual aesthetics of web sites. International Journal of Human-Computer Studies, 60, 269-298.
- Lewis, J.R. (1991) Psychometric Evaluation of an After-Scenario Questionnaire for Computer Usability Studies: The ASQ. SIGCHI Bulletin, 23 (1), 78-81.
- Lewis, J.R. (2002) Psychometric evaluation of the PSSUQ using data from five years of usability studies. International Journal of Human-Computer Interaction, 14(3&4), 463-488.
- Lidwell, W., Holden, K., & Butler, J. (2003). Universal Principles of Design. Gloucester, MA: Rockport.
- Lindgaard, G. & Dudek, C. (2003). What is this evasive beast we call user satisfaction? Interacting with Computers, 15, 429-452.

- Lloyd, B.H. & Gressard, C. (1984). Reliability and factorial validity of computer attitude scales. Educational and Psychological Measurement, 44, 501-505.
- Lund, A.M. (2001) Measuring usability with the USE questionnaire. STC Usability SIG Newsletter,8 (2). Retrieved from [http://www.stcsig.org/usability/newsletter/0110\\_measuring\\_with\\_use.html](http://www.stcsig.org/usability/newsletter/0110_measuring_with_use.html)
- Mahlke, S., Minge, M., & Thüring, M. (2006). Measuring multiple components of emotions in interactive contexts. In Proceedings of CHI 2006. Montreal, Quebec, Canada. ACM.
- 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.
- Marcoulides, G.A. (1989). Measuring computer anxiety: The computer anxiety scale. Educational and Psychological Measurement, 49, 733-739.
- Matthews, G., Jones, D.M. & Chamberlain, A.G. (1990). Refining the measurement of mood: the UWIST mood adjective checklist. British Journal of Psychology, 81, 17-42.
- Mehrabian, A. & Russell, J.A. (1974). An approach to environmental psychology. Cambridge, MA: MIT Press
- Nagamachi, M. (2002). Kansei engineering as a powerful consumer-oriented technology for product development. Applied Ergonomics, 33, 289-294.
- Nielsen, J. (1993). Usability Engineering. London: Academic Press.
- Nowicki, S. & Duke, M. P. (1994). Individual differences in the nonverbal communication of affect: The Diagnostic Analysis of Nonverbal Accuracy Scale. Journal of Nonverbal Behavior. Vol 18(1), 9-35.
- Nunnally, J.C. & Bernstein, I.H. (1994). Psychometric Theory. Third Edition. New York: McGraw-Hill.
- O'Donnell, C.R. & Eggemeier, F.T. (1986). Workload assessment methodology. In K.R. Boff, L. Kaufman, & J.P. Thomas (Eds.), Handbook of Perception and Human Performance. Vol. II. New York: Wiley.
- Osgood, C.E. (1969). The nature and measurement of meaning. In C.E. Osgood & J.G. Snider (Eds.), Semantic Differential Technique: A Sourcebook. Chicago: Aldine.
- Park, S., Choi, D., & Kim, J. (2004). Critical factors for the aesthetic fidelity of web pages: Empirical studies with professional web designers and users. Interacting with Computers, 16, 351-376.

- Paviou, P.A. & Fygenon, M. (2006). Understanding and predicting electronic commerce adoption : An extension of the theory of planned behaviour. MIS Quarterly, *30*, 115-143.
- Petty, R.E., Fabrigar, L.R., & Wegener, D.T. (2003). Emotional factors in attitudes and persuasion. In R.J. Davidson, K.R. Scherer, & H.H. Goldsmith (Eds.), Handbook of Affective Sciences. Oxford: Oxford University Press.
- Picard, R.W. (1997). Affective Computing. Cambridge, MA: MIT Press.
- Power, M.J. (2006). The structure of emotion: An empirical comparison of six models. Cognition and Emotion, *20*, 5, 694-713.
- Raub, A.C. (1981). Correlates of computer anxiety in college students. Unpublished PhD. Thesis. University of Pennsylvania.
- Reid, D. (2004). A model of playfulness and flow in virtual reality interactions. Presence, *13*, 451-462.
- Rosenthal, R., Hall, J.A., DiMatteo, M.R., Rogers, P.L. & Archer D. (1979). Sensitivity to Nonverbal Communication: The PONS Test. Baltimore, Johns Hopkins University Press.
- Russell, J.A. (2003). Core affect and the psychological construction of emotion. Psychological Review, *110*, 145-172.
- Russell, J.A., Weiss, A. & Mendelsohn, G.A. (1989). Affect Grid: A single item scale of pleasure and arousal. Journal of Personality and Social Psychology, *57*, 3, 493-502.
- 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.
- Scherer, K. R. (2005). What are emotions? And how can they be measured? Social Science Information, *44*, 4, 693-727.
- Schütte, S.T.W, Eklund, J., Axelsson, J.R.C., & Nagamachi, N. (2004). Concepts, methods and tools in Kansei Engineering. Theoretical Issues in Ergonomics Science, *5*, 214-231.
- Sillence, E., Briggs, P., Peter, H., & Fishwick, L. (2006). A framework for understanding trust factors in web-based health advice. International Journal of Human-Computer Studies, *64*, 697-713.
- Sloman, A. (2002). How many separately evolved emotional beasies live within us? In R. Trappl, P. Petta, & S. Payr (Eds.), Emotions in Humans and Artefacts. Cambridge, MA: MIT Press.
- Smith, B., Caputi, P., & Rawthorne, P. (2007). The development of a measure of subjective computer experience. Computers in Human Behavior, *23*, 127-145.

- Spielberger, C.D., Gorsuch, R.L., Lushene, R., Vagg, P.R., & Jacobs, G.A. (1983). Manual for the State-Trait Anxiety Inventory (Self Evaluation Questionnaire). Palo Alto, Ca.: Consulting Psychologist Press.
- Tractinsky, N., Katz, A.S., & Ikar, D. (2000). What is beautiful is usable. Interacting with Computers, 13, 127-145.
- Van der Heijden, H. Verhagen, T., & Creemers, M. (2003). European Journal of Information Systems, 12, 41-48.
- Voss, K.E., Spangenberg, E.R., & Grohmann, B. (2003). Measuring the hedonic and utilitarian dimensions of consumer attitude. Journal of Marketing Research, 40, 310-320.
- Vrendenburg, K., Mao, J-Y., Smith, P.W., & Carey, T. (2002). A survey of user-centred design practice. CHI Letters, 4, 1, 471-478.
- Wang, Y-S. Tang, T-I., Tang, J-t E. (2001). An instrument for measuring customer satisfaction toward web sites that market digital products and services. Journal of Electronic Commerce Research, 2, 89-102.
- Watson, D., Clark, L.A. & Tellegen, A. (1998). Development and validation of brief measures of positive and negative affect: The PANAS scales. Journal of Personality and Social Psychology, 54, 1063-1070.
- Webster, J., Trevino, L.K., & Ryan, L. (1993). The dimensionality and correlates of flow in human-computer interactions. Computers in Human Behavior, 9, 411-426.
- Westerman, S.J., Shaerf, S., Tuck, G.C., & Gardner, P.H. (manuscript under review). Assessing the structure of users' responses to website designs.
- Youngstrom, E.A. & Green, K.W. (2003). Reliability generalization of self-report of emotions when using the Differential Emotions Scale. Educational and Psychological Measurement, 63(2), 279-295.
- 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.