On the moment-to-moment measurement of emotion

On the moment-to-moment measurement of emotion during person-product interaction

Gaël Laurans

Gaël Laurans

On the moment-to-moment measurement of emotion during person-product interaction

Gaël Laurans

http://www.laurans.ch/

On the moment-to-moment measurement of emotion during person-product interaction

by means of video-supported retrospective self-report, with some ancillary remarks on other issues in design-related emotion measurement

Proefschrift

ter verkrijging van de graad van doctor aan de Technische Universiteit Delft, op gezag van de Rector Magnificus prof. ir. K.C.A.M. Luyben, voorzitter van het College voor Promoties, in het openbaar te verdedigen op donderdag 22 december 2011 om 15.00 uur

door Gaël François Gérard LAURANS

DESS Psychologie du travail et nouvelles technologies, Université de Metz en Université Nancy 2, geboren te Saint-Julien-en-Genevois, Frankrijk.

Dit proefschrift is goedgekeurd door de promotor:

Prof dr. P.P.M. Hekkert, Technische Universiteit Delft

Copromotor: Dr. ir. P.M.A Desmet

Samenstelling promotiecommissie:

Rector Magnificus, voorzitter
Prof. dr. P.P.M. Hekkert, Technische Universiteit Delft, promotor
Dr. ir. P.M.A. Desmet, Technische Universiteit Delft, copromotor
Prof. dr. J. Schoormans, Technische Universiteit Delft
Prof. Dr. G. Cupchik, University of Toronto
Prof. dr. P. van Schaik, Teesside University
Prof. dr. M. Neerincx, Technische Universiteit Delft
Dr. W. IJsselsteijn, Technische Universiteit Eindhoven
Prof. dr. R. Huib de Ridder, Technische Universiteit Delft, reservelid

Table of contents

1. Introduction

Emotion is now firmly established as a major focus in product design and human-computer interaction. Over the last 10 years, research on design and emotion has flourished. Conferences on the topic are organized regularly and two series are dedicated exclusively to the topic. The first *Design & Emotion* conference started in Delft in 1999 and grew from an event with 41 participants to a multi-track conference with hundreds of participants. The last edition (2010 in Chicago) had over 250 communications and the next one is already scheduled for 2012 in London. *Designing Pleasurable Product & Interfaces* is another series of events devoted to the affective side of product design and human-computer interaction. Its first iteration dates from 2003 in Pittsburgh and the 5th edition was organized in Milan in 2011. The publication of several influential monographs (Jordan, 2000; Norman, 2004) and collections of articles (Blythe, Overbeeke, Monk, & Wright, 2003; McDonagh, Hekkert, Erp, & Gyi, 2003) further illustrates the development of the field.

Affective processes and experiences have also been identified as important phenomena in related disciplines such as consumer psychology and human-computer interaction with the emergence of the field of affective computing (Picard, 2010) and a renewal of interest for emotions in advertising (Poels & Dewitte, 2006), food science (King & Meiselman, 2010), and consumer research (Richins, 1997). Major human-computer interaction conferences like the Association for Computer Machinery's CHI also open considerable space to user experience (e.g. Law, Roto, Hassenzahl, Vermeeren & Kort, 2009).

Business writers have also popularized the idea that pleasure and affect are playing an increasing role in the marketplace, coining expressions like "experience economy" or "dream society" (e.g. Jensen, 1999). They explain that advanced technology, extra functionality, reliability and performance are not enough to satisfy customers anymore. To get an edge over their competitors, companies need something more than well-functioning products and offer designs their users can enjoy beyond pure utility.

1.1. The Science of Emotion

At the same time, research on emotion has seen a resurgence within psychology, starting in the 1970s and culminating in the creation of new journals (e.g. *Emotion* in 2001; *Emotion Review* in 2009) and fundamental texts like the *Handbook of Emotions* (1st edition 1993, 3rd edition 2008), the *Handbook of Cognition and Emotion* (1999) and the *Handbook of Affective Sciences* (1st edition 2003, 2nd edition 2009). Neuroscience has also increasingly looked at affective processes as illustrated among others by Antonio Damasio's famous 1994 book, *Descartes' Error.*

Researchers in these fields criticize what they see as the traditional understanding of affect as an uncontrollable, subjective phenomenon that is inaccessible to scientific study and emphasize the evolutionary role of emotions. Far from being a dysfunctional process that disturbs rational thinking and only produce maladaptive behavior, emotions help us to quickly face challenges and seize opportunities in our environment. Affect is therefore a mechanism that allows us to rapidly evaluate what is happening around us and react appropriately without relying solely on slow and costly deliberate thinking. Positive emotions motivate us to seek beneficial situations and outcomes but also to engage with the world, fostering exploration, creative problem solving, and long-term well-being (Fredrickson, 2001).

1.2. Implications for Design

All these effects underline the importance of emotion for design, as it is a major force directing our behavior, including buying or using products in everyday life. Thus emotions are much more than the proverbial icing on the cake; they are an integral part of any interaction with the world and contribute to the myriad of decisions we make about choosing, adopting, using, retaining, recommending or abandoning products.

However, the emotions that can be expected in relation to products are likely to be somewhat different than the affective states found in current psychological research. For example, responses to product design are often milder than the feelings experienced in interpersonal situations. Design research is also likely to be more interested in subtle positive experiences than the strong negative responses studied in clinical psychology.

A number of researchers have been looking for ways to provide designers with insights and approaches to deal with these emotions in their work. This thesis is more specifically devoted to techniques to

assess the emotions we experience as we use and *interact* with products. The goal is to contribute to the development of measurement procedures that can be used in design-oriented research to better understand the role of emotions in interaction between several kinds of products and their users.

It differs from other similar efforts (Desmet, 2002; Karapanos, 2010; Russo, 2010) by its focus on short episodes of interaction. Desmet studied people's response to the appearance of products presented to them statically (i.e. as pictures or simply displayed on a table) with a questionnaire designed for this purpose. Karapanos and Russo also devised their own measures to look at different aspects of product experience but focused on long-term relationships (how one's attitudes and feelings toward a product change and develop over months or years).

1.3. The Present Thesis: Emotion and interaction

By contrast, the present work is focused on immediate changes in feelings following a sequence of interaction with a product and on the dynamics of experience over minutes and hours. Understanding these short-term changes in experience is becoming increasingly important with the multiplication of programmable interactive products. For example, using a personal navigation device or other in-car systems involves multiple elementary actions spread over time, and designers do not only create the physical shape of the device or a few isolated mechanisms like changing the memory card but also need to define the response of the system during complex sequences of interaction (looking for alternative routes, integrating external information about traffic, etc.)

This object of study raises particular challenges that have rarely, if ever, been addressed directly in the scientific literature, whether fundamental (i.e. psychology) or applied (including media or consumer psychology, human-computer interaction and design research), in particular the need to collect moment-to-moment measures of mild affective responses while research participants are busy with using a product or device and unavailable to report their feelings.

1.4. Structure of the Thesis

Chapter 2 presents some aspects of emotion and provides an overview of the various approaches available to measure them, discussing their usefulness for the evaluation of responses to product design in general

and person-product interaction in particular. The review covers both punctual (i.e. after the fact) and moment-to-moment tracking of the dynamics of experience.

While many promising methods were identified in chapter 2, few if any of them have been used in published studies of person-product interaction. Chapter 3 describes two such studies, using well-known questionnaires to collect punctual ratings of emotional experience after short sequences of interaction with different products (coffee machine, alarm clock, personal navigation devices).

Chapter 4 turns to moment-to-moment measurement and details the challenges faced by researchers interested in the dynamics of experience. It sketches an approach to deal with them and adapt methods from other fields to this particular context. Chapter 5 presents empirical research on self-confrontation (video-supported retrospective interview), a major component of this approach. It details two studies that represent the first attempts at extending self-confrontation to affective phenomena in person-product interaction and to integrate it with quantitative approaches to moment-to-moment changes in affect.

Chapter 6 describes the design and empirical evaluation of the emotion slider, a device conceived to facilitate self-report during the self-confrontation procedure. A series of experiments with static pictorial stimuli was conducted to better understand the characteristics of the device before using it to collect moment-to-moment ratings of affective experience.

Chapter 7 and 8 discuss several issues related to the reliability and validity of measures of emotion, including both short-term moment-to-moment and design-oriented research in general. The conclusion (chapter 9) briefly evokes implications for design and some perspective for future research.

2. Measuring Affect

The sheer number and variety of instruments used to measure affect is impressive[1]. Numerous quantitative studies of emotion have appeared in social psychology but also in fields like design, advertisement or media psychology, and human-computer interaction. Despite this broad interest in emotion, measures are rarely standardized and studies on their psychometric qualities (validity, reliability) are still relatively rare.

Empirical studies often rely on *ad hoc* single-item scales or measurement techniques chosen for convenience and most multi-item questionnaires found in the literature have been developed with clinical research in mind. Other approaches such as physiological measurement have also primarily been developed and tested with strong clinically relevant affective stimuli and are rarely examined from a psychometric perspective. All this makes a comparison between measures and an evaluation of their appropriateness for design-oriented research particularly arduous.

Additionally, the emotions that can be expected during product-person interactions differ in several ways from those experienced during major life events or laboratory studies. Products typically elicit mild and subtle responses rather than intense full-fledged prototypical emotions. Products are also more complex and ambiguous than many stimuli used in psychological research.

Other fields, such as consumer psychology, advertisement research, human computer-interaction, affective computing, software and web usability, media psychology, and music perception face similar issues and many relevant empirical studies have been published, dating back at least to the 1980s. Often, however, these studies simply adapt methods from basic or clinical research, ignoring work from neighboring fields, and the results are then promptly forgotten until a new questionnaire or a new technique comes along.

This review will organize this scattered literature following a multi-

1 Following widespread usage in emotion psychology (Ekman & Davidson, 1994; Russell, 2003), affect is understood here as a general label for a number of related phenomena including moods (long-lasting, diffuse affective states) and emotion *sensu stricto* (brief, conscious affective responses to a specific object or event). Consequently, "affect" encompasses both moods and emotions. It will also occasionally be used, especially in the adjective form ("affective"), to avoid constantly repeating the word "emotion" where the distinction is not essential and the context precludes any ambiguity.

componential view of emotion (Scherer, 2005) associating each measurement tool to one of the main facets of emotion: conscious feeling, bodily changes, expression and behavior. This organization also makes it possible to relate measurement problems to salient aspects of the psychological literature on these components. Chapter 8 will build upon this review to dispel widespread confusions about the validity of different types of measures of affect.

Finally, the relevance of each component to the moment-to-moment measurement of on-going emotional responses will be assessed. This assessment forms the basis of the development of the measurement procedure described in chapter 4 and 5.

2.1. Feelings/self-report

Feelings – the conscious experience of the emotion itself – are a key component of emotion. Even if current research emphasizes unconscious affective processes, feelings still form the core of our intuitive understanding of emotion and the starting point for investigations into other components. As such, self-report enjoys a high face validity which, combined with its ease of use and versatility, has made it the most common family of emotion measures.

2.1.1. Self-report scales

Self-report instruments can be divided in two groups depending on the form of the items: verbal tools use words or sentences to describe feelings whereas graphical tools are based on depictions of emotions with cartoon faces or animated characters. In all cases, research participants are asked to choose the words or pictures that best match their current state or to rate how close each item is to their feelings. Open-ended questioning or text mining can also be linked to self-report as they rely on people's verbalization of their conscious experience.

Adjective checklists or rating scales are certainly the most common self-report instruments and the POMS (profile of mood scales) and the MAACL (multiple affect adjective checklist) probably the most successful of several similar checklists developed in the 1960s and 1970s.

The POMS was published in 1971, with several revisions, a new bipolar version (Lorr, 1989) and several short forms released later. The traditional version includes 65 items organized in six dimensions (anger/hostility, depression/dejection, vigor/activity, fatigue/inertia, confusion/bewilderment, tension/activity). Participants have to rate their current state on a five-point response format ("not at all" to

"extremely"). POMS-BI, the bipolar version, uses 72 adjectives with a different response format (four points: "much unlike this", "slightly unlike this", "slightly like this", "much like this") and six bipolar dimensions (composed–anxious, agreeable–hostile, elated–depressed, confident–unsure, energetic–tired, clearheaded–confused).

Unlike the POMS, the MAACL is a pure adjective checklist. Respondents are simply asked to select which words fit their current state in a list. The first version, based on several earlier instruments, was published in 1965 (Zuckerman, Lubin & Rinck, 1983). A revised version, the MAACL-R, was developed in the 1980s (Zuckerman et al., 1983; Zuckerman et al., 1986), correcting some of the issues that emerged with the older scales (Gotlib & Meyer, 1986; Thayer & Sinclair, 1987; Zuckerman et al., 1983) while keeping the adjective checklist format. This revised version has 132 adjectives, grouped in three bipolar negative scales (anxiety, depression and hostility) and two unipolar scales (general positive affect and sensation seeking).

Robert Plutchik developed several questionnaires based on his psychoevolutionary theory of emotions. It is centered on eight primary emotions, which can have different names depending on the "language" or level considered. Thus protection, destruction, reproduction, reintegration, incorporation, rejection, exploration and orientation ("functional language") can respectively be called fear, anger, joy, sadness, acceptance, disgust, expectancy and surprise in the "subjective language". Each of these emotions corresponds to a basic adaptive need and can be combined to describe all other emotions (for example love is a composite of joy and acceptance). Plutchik developed several self-report instruments to measure the primary emotions, which led to some confusion in the literature. The most important one is probably the Emotions Profile Index (EPI; Kellerman & Plutchik, 1968), a questionnaire based on forced choice between 62 or 66 combinations of 12 personality traits (i.e. for each pair, the participants have to indicate which one is more like themselves). Each of these traits is associated with two of the eight primary emotions, allowing the researcher to build an "emotion profile" for each participant. The EPI was developed for patients in a psychiatric hospital and has been used mostly in clinical psychology. Another instrument, the Emotion-Mood Index is a more traditional adjective checklist with 72 items grouped in nine clusters or dimensions (the eight primary emotions plus an arousal cluster, see Plutchik, 1980). Plutchik (1966, 1980) also used various brief rating scales with only one adjective for each primary emotion.

Another influential framework is Caroll Izard's differential emotion theory (Izard, 1971). This theory postulates nine fundamental emotions (although Izard himself occasionally stressed that his list was not thought to be definitive): interest, joy, surprise, distress, anger, disgust, contempt, shame and fear. Each emotion is thought to be

associated with different patterns of neural activity, facial-postural activity and subjective experience. The Differential Emotions Scale (DES) is a self-report instrument based on this theory. The first version was developed by selecting common adjectives used by participants to label facial expressions for each of the fundamental emotions. The scales were then refined and reduced to three adjectives per scale based on factor analyses of current mood ratings by two student cohorts. Several studies tested the validity of the DES by looking at self-reported mood in various situations.

Mehrabian's Pleasure-Arousal-Dominance (PAD) scales are a very different set of adjective rating scales. Unlike the various questionnaires discussed above, PAD is not designed to measure discrete emotions but three broad dimensions of affect. Russell & Mehrabian (1977) proposed that other scales and specific emotions can be mapped to the space defined by these dimensions and that pleasantness, arousal, and dominance provide the most economical description of emotions. Mehrabian (1996) also suggested that these three dimensions underlie personality and various types of cognitive judgments. The first version of the PAD questionnaire was composed of 18 pairs of opposite adjectives with a 9-point response grid (Mehrabian & Russell, 1974). Respondents have to describe their current state by ticking a box between each pair of adjectives. Other versions with a different number of items but with the same general structure exist (Mehrabian, 1995).

Russell's Affect Grid was designed to quickly assess the first two PAD dimensions, namely valence (pleasure) and arousal, with a single item in the form of a 9 x 9 grid, anchored by 8 words spread around it (Russell, Weiss & Mendelsohn, 1989). Respondents have to indicate their current state by checking one of the boxes in the grid.

The positive and negative affect schedule (PANAS) is a 20-item adjective-rating instrument presented in Watson, Clark & Tellegen (1988). The questionnaire is made of two 10-item scales, measuring positive and negative affect. Participants have to indicate how well words like "interested", "distressed" or "nervous" describe their affective state on a scale from 1 ("very slightly or not at all") to 5 ("extremely"). Large-scale studies (Crawford & Henry, 2004; Crocker, 1997; Mackinnon et al., 1999; Watson & Clark, 1994) have found support for the bidimensional structure of the questionnaire but also a small negative correlation between both scales. An expanded version of the PANAS (the PANAS-X) is also available, adding 11 lower order specific affect scales (fear, hostility, guilt, sadness, joviality, self-assurance, attentiveness, shyness, fatigue, serenity and surprise) to the two general dimensions, for a total of 60 items. Interestingly, the relevant PANAS-X subscales (fear, hostility, sadness, fatigue and positive affect) seem to be highly correlated with the POMS scales (tension-anxiety, anger-hostility, depression-dejection, fatigue, vigor),

while having generally lower interscale correlations.

Thompson, E.R. (2007) developed an abbreviated version of the PANAS (called I-PANAS-SF) specifically designed for proficient but non-native speakers of English (for example students at internationally oriented universities or employees in transnational corporations). Besides being briefer while retaining adequate content coverage and psychometric qualities, I-PANAS-SF also avoids several items that proved difficult in previous studies: "jittery" (Laurans, 2009; Thompson), "excited" (Dubé & Morgan, 1996; MacKinnon et al., 1999), and "distressed" (Laurans; Thompson).

The Evaluative Space Grid (Larsen, Norris, McGraw, Hawkley & Cacioppo, 2009) is a single item instrument structurally similar to the affect grid but based on the same dimensions as the PANAS. Instead of pleasure and arousal, one axis reflects the amount of positive feelings and the other the amount of negative feelings, with instruction stressing that positive and negative feelings can also co-occur.

The Self-Assessment Manikin (SAM), the most common non-verbal self-report instrument, is another tool derived from PAD. Bradley & Lang (1994) report a validation study comparing the non-verbal SAM to the verbal PAD scales. Instead of pairs of adjectives, each dimension is pictured by a series of five schematic characters. For example, varying the shape of the mouth from a frown to a large smile represents different degrees of pleasure and displeasure. Since the drawing themselves are quite abstract and the precise meaning of the different dimensions can be difficult to grasp, use of the SAM is usually preceded by extensive verbal instructions, anchoring each scale with a range of adjectives. Because a single graphical item replaces each 6-item scale, SAM is much quicker to administer and has been extensively used, in particular to standardize sets of affective stimuli (Bradley & Lang, 2007).

PrEmo (Desmet, 2002) is another graphical feeling questionnaire. Using animated cartoons to represent a set of emotions, it is the only purely non-verbal feelings self-report tool. People are known to attribute emotions to facial configurations (Matsumoto, Keltner, Shiota, O'Sullivan & Frank, 2008; Russell, Bachorowski & Fernández-Dols, 2003), body position (Wallbott, 1998) or movements (Bassili, 1978, 1979; Visch & Goudbeek, 2009). Dynamic facial expressions have also been shown to induce clearer mimicry than static displays (Sato, Fujimura & Suzuki, 2008). PrEmo's cartoons take advantage of all these effects to display more expressive depictions of each emotion. Combining animation and sound allows portraying these emotions without using any affective words, even in the instructions.

In practice PrEmo is administered on a computer: research participants click on each of the character in turn and, after seeing the animation, can register their rating to indicate the extent to which they experience the corresponding emotion with a three ("not at all",

"a little", "a lot") or five points response format.

While 14 animations are available in total in the current version of PrEmo, most studies use only the 10 most commonly reported emotions. This standard set includes five positive (desire, amusement, satisfaction, fascination, pleasant surprise) and five negative emotions (contempt, disgust, dissatisfaction, boredom, unpleasant surprise), originally selected for their relevance to product design (Desmet, 2004).

2.1.2. Use in applied research

Most of the instruments described above (PANAS, POMS…) were originally conceived as mood measures, assessing a diffuse affective state rather than a brief response to a particular event or situation. The main exception is obviously PrEmo, as it was developed specifically to assess emotions associated with products.

Published studies using PrEmo include research on car appearance (Desmet, 2004; Desmet, Hekkert & Hillen, 2004; Desmet, Hekkert & Jacobs, 2000), mobile phones (Desmet, Pocelijn & Van Dijk, 2007) and wheelchairs for children (Desmet & Dijkhuis, 2003).

Mood questionnaires can however also be used to measure the effect of a product, in a before-after design or by comparing reports obtained after using different products.

For example, Dubé & Morgan (1996) studied patients' experience of a hospital stay and Mooradian & Olver (1997) conducted a survey of peoples' feelings about their current car with the PANAS. Huang (1997) used it to investigate different models of the effect of negative affect on persuasion and attitude toward ads but recommended the use of discrete scales in her conclusions.

Plutchik's work has also had some influence on marketing research but despite frequent references to his theory in general and to the Emotion Profile Index in particular across the advertisement and consumer experience literatures, none of his measurement instruments seem to have been used in actual empirical research in these fields. Morris Holbrook (Havlena & Holbrook, 1986; Holbrook & Westwood, 1989) did however develop his own measure of Plutchik's primary emotions, using *a priori* scales with three adjectives for each emotion. Zeitlin & Westwood (1986) also describe a similar set of self-report scales but do not provide much information on the characteristics of the instrument.

Westbrook & Oliver (1991) used the DES in a study with owners of newly purchased cars. They were able to show that two different patterns of emotions can lead to high satisfaction.

2.1.3. Interpretation issues

The most thoroughly discussed question regarding affective self-report data is the list or model of emotions needed to properly represent affective experience. Many questionnaires include a relatively high number of scales, conceived as measures of separate, discrete emotions. They are often interpreted as basic emotions, i.e. innate responses to different evolutionary challenges or fundamental processes underlying common psychiatric diagnoses. The main alternative to this discrete emotions approach are dimensional models of emotion, based on a limited number of broad dimensions such as valence or arousal.

In recent years, "basic emotions" models have been mostly associated with research on facial expression (Izard, 1971; Ekman, 1999) and dimensional models with different types of self-report (Barrett & Russell, 1999; Watson, Clark & Tellegen, 1988) but both have been applied to all kinds of data. In fact, many clinical self-report questionnaires or affective checklists (Lorr, 1989; Nowlis, 1965; Zuckerman et al., 1983) attempt to measure – mostly negative – discrete emotions. The list and names of the emotions included vary but they usually include at least sadness/depression/distress, anger/hostility and fear/anxiety. Ekman's (1992) influential list of basic emotions (happiness, fear, disgust, surprise, anger, sadness, surprise) has not been turned into a systematic self-report instrument but Izard's (1971) DES and Power's (2006) Basic Emotions Scale draw on similar sources and assess almost the same emotions (omitting surprise for Power and adding a few other emotions – interest, shame, shyness, guilt and contempt – for Izard).

However these questionnaires suffer from several empirical problems, including difficulties to recover the hypothetical subscales in factor or component analyses of self-report data and lack of divergent validity between these subscales. Indeed, different negative subscales tend to be highly correlated, lending support to the notion that emotions are organized along a small number of broad dimensions and that self-report questionnaires mostly measure indiscriminate positive or negative affect. Studies on advertisement (Holbrook & Westwood, 1989) and consumption experience (Havlena & Holbrook, 1986) also suggest that discrete emotion indices based on Plutchik's theory did not add information compared to a tridimensional questionnaire.

Such findings support the notion that between one and three dimensions can account for the bulk of the variance in self-report of affect. Such models have a long history in psychology, with many researchers focusing either on pleasure or arousal alone (Yik, Russell & Barrett, 2009). One influential model postulates that pleasure (or valence) and arousal (or activation) defines a two-dimensional space

summarizing momentary affective experience. Specific emotions or ambivalent feelings then result from rapid changes in feelings or the combination of this "core affect" with other processes of a more cognitive nature (Russell, 2003). More specifically, James Russell has long insisted (Barrett & Russell, 1999) that the two fundamental dimensions of affect are bipolar and that affective experiences or the words describing them are not evenly spread in the whole space, instead forming a circle or circumplex within that space (Russell, 1980).

Another influential dimensional model is David Watson and Auke Tellegen's positive activation/negative activation framework (Watson, Wiese, Vaidya & Tellegen, 1990). While emphasizing the broad agreement between the different circumplex models, they argue that it is often more convenient to describe the affective space using two unipolar dimensions: positive and negative activation. Positive activation is associated with a general approach system and also with extraversion measures in personality inventories. Conversely, negative activation is associated with an avoidance or withdrawal system and with neuroticism. While they are based on two distinct biological systems, self-report ratings of positive and negative activation are often negatively correlated. Emotion data can therefore be analyzed as a three-level hierarchy (Tellegen, Watson & Clark, 1999; Watson, Wiese & al.). At the lowest level of the hierarchy, discrete emotions like those measured by the Differential Emotions Scale are clearly distinguished by factor analysis but also correlate with each other. At the next level in the hierarchy, two second-order factors, positive and negative activation, can be identified. Finally, the bipolar valence (pleasantness-unpleasantness) dimension can be extracted as an overarching third-order factor.

This hierarchical model can therefore reconcile the idea that a single dimension is not enough to give a full description of affective states (Barrett & Russell, 1999; Fontaine, Scherer, Roesch & Ellsworth, 2007; Larsen, Norris, McGraw, Hawkley & Cacioppo, 2009) with the finding that valence or pleasantness accounts for a big part of the variance in emotion data and could form a basic building block for emotion theory (Barrett, 2006).

Another important issue with many emotion measurement questionnaires described in the literature is their almost exclusive focus on negative affect. Clinical scales often include a single undifferentiated "positive affect" scale, sometimes two (typically joy/satisfaction and interest). This limitation, already noted by emotion researchers (Lorr & Wunderlich, 1988; Zuckerman & Lubin, 1990; Zuckerman et al., 1983) has been identified as a key problem for applied use (Desmet, 2002; King & Meiselman, 2010). For example, the distinction between anxiety, hostility and depression – the main focus of empirical research on these questionnaires in psychopathology – does not seem very

relevant for design-related research and Zuckerman et al. observed that most participants outside of clinical samples report extremely low scores on MAACL scales for these negative emotions. Holbrook and Westwood (1989) and Havlena and Holbrook (1986) also found high correlation between different negative emotion indices and a general measure of (dis)pleasure, further undermining the empirical relevance of the distinction between them for consumer research.

2.2. Bodily arousal

Another major component included in componential models of emotion is bodily arousal, i.e. all the changes in the inner organs (heart, viscera…) commonly experienced with emotions. Historically, the measurement of these changes and the study of their impact on affective processes is the main research topic in psychophysiology.

Psychophysiological research studies many signals, some of them more common than other for a number of reasons. Often, the choice of signals to record depended just as much on practical convenience as on theoretical soundness (Kreibig, 2010). This review is organized by response system, grouping measures reflecting activity in a set of functionally related organs (e.g. the cardiovascular system includes the heart, arteries, veins and capillaries). Each response system influences several signals, collected with different sensors. Only the most common systems and a few less common ones that have been considered in applied fields (affective computing and human-computer interaction) are described here.

"Wet" or neuroendocrine psychophysiology is the part of psychophysiology concerned with changes in the hormonal composition of the blood. These techniques can be very informative, especially in the context of stress research, but they are very intrusive and therefore seldom used outside of medical research. Electrophysiology ("dry" psychophysiology) is based on the measurement of different kind of electrical signals resulting from the functioning of the body, especially neuron firing.

Electrophysiological techniques are used to study the autonomic nervous systems (e.g. electrocardiography, electrodermal activity), muscle activity (through electromyography) or brain activity (electroencephalography). Only the first set of measurement will be discussed in this section. Electromyography and electroencephalography are very similar to electrocardiography on a technical level but they tap into completely different neural processes and response systems and will be discussed in section 2.3.

2.2.1. Response systems and measurement techniques

Electrodermal activity (EDA) includes all changes of the conductance of the skin under the influence of minute differences in sweating activity. It is the most frequent measure in research on the psychophysiology of emotion but generally lacks specificity. Increased electrodermal activity accompanies all emotions except certain forms of sadness, contentment and relief, suggesting it is related to motor preparation – affective or not (Kreibig, 2010). Beside its role in thermoregulation, sweating has also been shown to be related to a number of psychological processes (see Boucsein, 1992, for a comprehensive reference; Fowles et al., 1981, for guidelines from a leading psychophysiology journal; Hugdahl, 1995, for a good overview). These conflicting influences threaten its validity as an emotion measure (see also interpretation issues below and chapter 8).

The cardiovascular system is another major response system studied in psychophysiology. It is regulated by several complex mechanisms, including endocrine and nervous influences (see Hugdahl, 1995, chapter 9-10 and Papillo & Shapiro, 1990, for an overview; Berntson et al., 1997; Jennings, et al., 1981, and Shapiro et al., 1996, for technical guidelines). Kreibig (2010) lists over 30 different measures, the most common being heart rate and systolic and diastolic blood pressure. Cacioppo, Berntson, Larsen, Poehlmann, and Ito's (2000) meta-analysis of 13 studies meeting stringent methodological criteria found that heart rate could differentiate between some emotions, especially between disgust and other emotions. Kreibig's more inclusive qualitative review of 134 studies suggests that heart rate is more specifically related to the passivity of the emotion, decreasing for passive states such as contentment or sadness and increasing with more active states – both negative and positive – such as anger, anxiety and joy.

A few less common physiological measures such as pupil size and face temperature have attracted some interest in affective computing/ human computer interaction research because of their practical advantages.

Early psychophysiological research with affective pictures suggested pupil size changes with emotion (Hess & Polt, 1960). Bradley, Miccoli, Escrig & Lang (2008) and Partala & Surakka (2003) observed pupil dilation for affective tones and pictures, both pleasant and unpleasant, and a high correlation between pupil size and arousal and amplitude of skin conductance response suggesting it is mainly related to emotional arousal.

A few studies have also linked face temperature and blood flow to the head – which can be unobtrusively measured with infrared

thermography – to autonomic activity in stress and affective situations (Merla & Romani, 2007; Puri, Olson, Pavlidis, Levine & Starren, 2005).

2.2.2. Use in applied research

In spite of the technical difficulties and often ambiguous results, publications with psychophysiological techniques are in fact quite common in the applied literature, especially in human-computer interaction and in marketing research.

Wang and Minor (2008) found 67 marketing-related studies including psychophysiological measures (not all emotion-related).

Jenkins, Brown, and Rutterford (2009) asked research participants to imagine preparing a hot drink using several products and found some relationship between infrared thermography of the face and electroencephalographic data. Puri et al. (2005) and Jenkins et al. suggest that the technique could be used to monitor stress and frustration or assess emotional state in human-computer interaction and design research.

Ward and Marsden (2003) and Westerman, Sutherland, Robinson, Powell, and Tuck (2007) both included psychophysiological signals in their measures of user responses to websites. Ward and Marsden asked their participants (N = 20) to find some information in two different websites (an "ill-designed" website and a "well-designed" one). They could not identify any significant difference between websites in the skin conductance, heart rate and finger blood pulse volume data. Westerman et al. asked their participants (N = 40) to passively browse two pages on two websites presented either in full color or in black and white. Only the color manipulation had an effect on skin conductance, with a lower skin conductance when the website was presented in black and white.

Mahlke, Minge, and Thüring (2006) and Mahlke and Thüring (2007) measured skin conductance and heart rate of participants using different on-screen prototypes of interactive products (audio player, mobile phone) and found some modest but significant correlations between self-report and physiological measures. Ravaja, Turpeinen, Saari, Puttonen, and Keltingas-Järvinen (2008) and Mandryk and Atkins (2007) also used skin conductance and heart rate in studies with video games.

2.2.3. Interpretation issues

While the psychophysiological literature documents many effects of emotion on bodily activity, these measures are particularly difficult to collect and interpret.

Most common physiological signals result from the integration of many complex processes and have been related to a host of phenomena beside emotions. For example, skin conductance responses can vary in amplitude depending on cognitive workload or the probability of an aversive event (Boucsein & Backs, 2000). The cardiovascular system also fulfils an important metabolic function and is obviously sensitive to physical activity. Changes in many physiological response systems have also been observed as part of the orienting response, an interruption of on-going processes following the apparition of any novel stimulus in the environment, including changes in light and sounds. Consequently, relationships between psychological events and simple physiological measures are typically many-to-one rather than one-to-one (Cacioppo & Tassinary, 1990).

For these reasons, physiological measures are generally very noisy and psychophysiological research typically requires a much more strictly controlled environment than research on other type of responses.

A more fundamental issue is the lack of invariance in physiological correlates of affective processes. Results in this field are subject to a great amount of interpersonal and contextual differences. For example, even when differences between stimuli are strong at an aggregate level, the correlation between the amplitude of the skin conductance response and self-reported arousal when viewing pictures might be non-significant for as many as 60% of the participants (Lang, Greenwald, Bradley & Hamm, 1993).

Discussing the results of a meta-analysis of psychophysiological studies on the differences between emotions, Cacioppo, Berntson, Larsen, Poehlmann, and Ito (2000, see also update in Larsen, Berntson, Poehlmann, Ito & Cacioppo, 2008) stress that results are contingent on the elicitation method. For example, a pattern of change associated with a given emotion might be observed when it results from imagery but not from hearing music or viewing pictures and vice versa.

There is also evidence that attempts to regulate or to hide emotions also have strong effects on bodily arousal (Gross & Levenson, 1997), further compounding the problem and calling into question the view of psychophysiological signals as objective measures isolated from participants conscious will (see also chapter 8).

2.3. Expressive behavior

Facial expression is probably the component of emotion that received the most attention in emotion research in the second half of the 20th century. Indeed, the study of facial expression has been ascribed a major role in renewing interest in emotions in general in a time when it was a neglected topic of research (Ekman 1993; Russell, Bachorowski

& Fernández-Dols, 2003).

Broadly speaking, two families of measurement techniques tap into facial expressions: observation and facial electromyography. In studies of human emotions, observation-based research usually employs elaborate coding systems and requires video or at least photographical recordings. Recently, facial observation has often been supplemented with computer-based classification of expressions to avoid time-intensive manual coding or even completely automate emotion recognition.

On practical and technical levels, facial electromyography (EMG) is quite different. It is in fact very similar to electrocardiography but instead of tracking heartbeats, it measures muscle activity with surface electrodes placed on the face. For this reason, it is often discussed together with the other psychophysiological techniques described above (e.g. Poels & Dewitte, 2006). Still, both observation of visible changes on the face and facial electromyography depends on activity of the same facial muscles and presumably on the same underlying brain systems and psychological processes.

Indeed, the neural circuits controlling facial muscles are very different from those controlling the cardiovascular system and the viscera. Heart function and blood circulation are regulated by the autonomic nervous system, especially through the spinal cord and vagus nerve, while facial muscles are skeletal muscles, mostly innervated by facial nerve VII (cranial nerve) and receiving influence from both pyramidal and extra-pyramidal (i.e. subcortical) pathways. We are also aware of our facial expressions and can to some extent control them deliberately (the level of control varies between regions of the face, see Rinn, 1984 for details).

2.3.1. Coding systems

Several coding systems have been developed to systematically assess facial movements based on video recordings. Ekman & Friesen's (original version 1978, newer electronic version: Ekman, Friesen & Hager, 2002) Facial Action Coding System (FACS) is an anatomically based comprehensive system that is not limited to affect displays. It can thus be used to represent any visible change on the face, without a priori theoretical assumptions on their relevance for the measurement of emotion. Facial movements are decomposed in elementary movements or "action units" (44 in the original 1978 version), which are the smallest units of movements that can be reliably detected by observers.

Since the FACS was explicitly developed to avoid any interpretation of the facial movements being coded, it does not directly produce any measure of emotion. However, FACS-based descriptions facial

expressions characteristic of various "basic emotions" have been published and the distributor of the FACS manual and training material also sells a subscription-based access to a database of FACS-coded expressions called the Facial Action Coding System Affect Interpretation Dictionary (FACSAID). These interpretation keys can be used to extra emotion measures from raw FACS-based description of facial movements. For example, low-level codes such as AU 4 + 5 ("brow lowerer" and "upper lid raiser") are interpreted as a sign of anger.

Several authors reported agreement between pairs of FACS raters between 73% and 85% (i.e. 85% of all movements in a given video have been coded identically by both raters and 15% are unique to one or the other rater). However, these agreement figures pertain to the movements coded and therefore cannot directly be interpreted as indicators of the reliability of FACS-based measurement of emotion[2].

The main practical hurdle to the routine use of the FACS is the time involved in the process. About 100 hours are needed to learn the FACS and the coding itself can take between one and three hours per minute of video (Bartlett, Hager, Ekman & Sejnowski, 1999; Matsumoto, Ekman & Fridlund, 1991). Several other approaches exist which allow a quicker analysis of facial movements when a full description is not needed.

EMFACS is a variant of the FACS relaxing some of the rules and limiting the coding to movements (action units) that proved relevant to the recognition of emotion in previous research. EMFACS is only available to trained FACS coders who passed the FACS final certification test but, according to its authors, it reduces coding time to about 10 minutes per minute of video.

Around the same time as Paul Ekman and Wallace Friesen were working on the FACS, Caroll Izard developed two facial expression coding systems, which have found some use mostly in developmental psychology (studies of infants and children). The Maximally discriminative facial movement coding system (MAX) is also based on facial anatomy and on the coding of elementary changes but it was streamlined to include only movements relevant to the measurement of fundamental emotions in Izard's differential emotion theory. Izard & Dougherty (1982) estimate the time needed to code a minute of video to vary between 20 and 200 minutes, which is somewhat less than the FACS but still much longer than many applied research settings

2 Interpretation keys often include several slightly different expressions for an emotion and many individual movements do not have any affective meaning. Consequently, disagreement between coders regarding the raw FACS codes does not automatically entail a disagreement on the emotional meaning of the overall expression. Conversely, a high level of agreement regarding irrelevant movements would not translate into high reliability of FACS-based measures of affect.

allow. Izard and Dougherty recommend using it in combination with another tool called the System for Identifying Affect Expression by Holistic Judgment (AFFEX). Unlike FACS or MAX, AFFEX is not based on the identification of elementary movement but on the evaluation of the whole expression by non-specialist judges. AFFEX provides a brief training procedure to improve the accuracy and reliability of these judges.

The Facial Expression Coding System (FACES) is a more recent system taking a similar approach as Izard's AFFEX, called by its authors the "cultural informant" approach. Untrained coders, supposed to be familiar with the culture of the person being filmed, are asked to provide judgment of the facial expression. FACES basically defines a set of instructions and a rating format to aid the non-expert coders to record their judgments. One of the key differences between this system and all the tools discussed above is the model of emotion underlying it. While FACSAID, EMFACS, MAX and AFFEX all attempt to measure a small number of basic or fundamental emotions (including typically fear, anger, disgust, etc.), FACES is based on a dimensional view of affect, asking judges to directly evaluate the valence and intensity of the expressions. It has been used in a number of studies, mostly in clinical psychology, and Kring and Sloan (2007) provide extensive evidence of convergence between FACES ratings of research participants watching emotion-eliciting videos and other measures (including EMFACS ratings, facial electromyography, autonomic physiology, and self-report of emotion). They also show that raters usually agree on the valence of the expressions.

FACEM (Katsikitis, Pilowsky & Innes, 1990; Pilowsky & Katsikitis, 1994) is another facial expression coding tool that received some use in clinical psychology. It combines simple manual coding and a computer model to make measurement as efficient as possible. Specifically, the coder must first identify the peak of a facial expression and then digitize 62 facial landmarks (80 in an earlier version) using a still picture and a graphics tablet. A model of the face is then used to automatically compute twelve distances and interpret them.

2.3.2. Automatic recognition of facial expression

Automatic recognition of facial expression has also been the focus of extensive research in affective computing, with the declared objective of making computer systems able to sense the affective state of their users without requiring any explicit input from them. Facial behavior is therefore a prime candidate for affective input as it can be monitored inconspicuously and continually with simple video equipment.

Facial expression recognition systems usually analyze photographs

or videos in several steps: first detecting the head and normalizing its position, then extracting facial features or landmarks from the pictures and finally feeding these schematized facial configurations to some machine learning algorithm to classify them in a few emotion categories. Before performing any effective recognition, machine learning systems need to be trained on a reference database containing pre-classified facial expressions. The accuracy and meaning of the results therefore depends on the quality of the training database and the way it was obtained.

Automatic recognition raises some new challenges of its own and a significant part of the research has understandably prioritized a range of technical issues including dealing with low-quality images, person-independence (recognizing expressions from persons not featured in the set of training pictures), choice of facial model and classifier (machine learning algorithm), and fusion between different modalities (e.g. information from the face and other signals) over directly addressing validity for applied research.

Most of the emotion recognition research has concentrated on the recognition of affective expressions from databases of posed facial behavior (Pantic, 2009), organized in six categories corresponding to Ekman's basic emotions (happiness, sadness, anger, fear, surprise, and disgust). It is difficult to provide an overview of classification accuracy, given the large differences between published studies in experimental design, stimuli used, model evaluation approach and indices of accuracy. Nevertheless, accuracies over 90% – i.e. on a set of pictures coded by humans, the system reports the same state (including neutral) as the human coders in 90% of the cases – have been reported in some conditions but the performance of systems trained on posed pictures is known to drop considerably when trying to classify real-life facial displays (Zeng, Pantic, Roisman & Huang, 2009). This difficulty is however a growing focus of current research in the field of affective computing and several studies about the automatic classification of naturalistic expressions have appeared (Pantic, 2009).

Another type of systems aims at recognizing elementary facial movements. Instead of producing a judgment about the emotion expressed, they output a set of FACS codes describing the expression itself (Bartlett et al., 1999; Cohn, Zochlower, Lien & Kanade, 1999). Automatic coding at the behavioral rather than emotional meaning level is particularly interesting for research, as it does not force researchers to trust a "black box" and to implicitly commit to interpretations of facial expressions that have been developed in other contexts. Such a system would make the identification and characterization of facial behavior occurring in applied settings much easier and enable research into its relevance for the measurement of design-related emotions.

Several research groups have been particularly active in the area

and adopted different approaches to automatic coding. The successive versions of the Automated Facial Image Analysis (AFA, see Cohn & Kanade, 2007) system developed at Carnegie Mellon University and at the Universiy of Pittsburgh are all based on the identification of several facial features (e.g. contour of the eyebrow, corners of the mouth) with local templates which are then used to detect FACS action units based on a-priori formulas (Cohn, Kanade, et al., 2001) or a classification algorithm (Cohn, Zochlower, et al., 1999). Michel Valstar and Maja Pantic (initially Delft University of Technology, now Imperial College London) developed another recognition system tracking 20 points on the face. Features describing the movement and distance between these points are then used to detect facial action units but unlike Carnegie Mellon's AFA, parameter selection for each classifier is entirely data-driven, not pre-constrained (Valstar & Pantic, 2006). The system developed at the University of California at San Diego's Machine Perception Lab uses filters to decompose the pictures and feeds the parameters to a learning algorithm without explicitly localizing any point or feature on the face (Bartlett et al., 1999). All these groups reported accuracies between 80 and 90% for their best algorithms when operating on sets of controlled posed expressions (Bartlett et al., 2006; Cohn, Zlochower, et al.; Valstar & Pantic, 2006), a performance similar to the level of agreement observed between expert coders[3]. Research with spontaneous data

3 The most common performance indicator is accuracy, i.e. percentage of agreement between the output of the recognition system and reference labels by expert FACS coders. These figures are somewhat comparable to the FACS inter-coder agreement but are only a partial description of the performance of an automatic coding system, which also depends on the set of choices in the test dataset and the prevalence of each expression in the situation of interest.

Accuracy is especially problematic when the classes have different sizes. When each expression is only present in a few pictures in the test set, overall accuracy will be mostly driven by classification efficiency for negative exemplars (i.e. neutral pictures and other expressions) and by the false alarm rate. The overall percentage of agreement with reference labels can be high even for a system with a low sensitivity (i.e. high false negative rate) because most pictures in the test set will be correctly categorized as *not* representing the particular action being tested.

When the test set is evenly balanced between positive and negative exemplars, accuracy will reflect both the sensitivity and false alarm rate but another counter-intuitive effect, often discussed as "base-rate neglect", might occur when using the system in a situation where the behavior of interest is rare: most of the cases flagged will be false alarms despite the good performance on the test set.

For example, both Bartlett et al. (2006) and Valstar and Pantic (2006) report an average accuracy above 90% in the recognition of many facial action units (20 AU for Bartlett et al., 15 for Valstar & Pantic). In the first case, the system was tested on a database including all expressions and many neutral pictures

(deception experiments, interviews) yielded more mixed results, with low hit rates for the recognition of 19 action units (Bartlett et al., 2006), some difficulties in categorizing movements in the brow area and some encouraging results in detecting blinks or smiles.

A practical problem faced by researchers willing to use automatic facial expression recognition in applied settings is that the various systems described in the literature are all experimental systems, sometimes available freely on the web or from their developers but difficult to deploy without considerable expertise. Ready-to-use software packages are however beginning to appear and to be applied to assess user's emotion during usability testing (Den Uyl & Van Kuilenburg, 2005). According to its developers, this particular system also performed well on the classification of elementary movements but this version is not commercialized (Den Uyl & Van Kuilenburg; Van Kuilenburg, Wiering & Uyl, 2005).

2.3.3. Facial electromyography

Following the renewed interest in facial expression, different researchers have shown that affective processes are associated with facial muscle activity measurable through electromyography (Cacioppo & Petty,

---

and the accuracy is high in spite of a low sensitivity (only 15% of AU are identified on average). In the second case, the test database is more balanced and the sensitivity is much better at 73% – the performance difference might result from the fact that Valstar and Pantic analyze whole sequences of posed facial behavior whereas Bartlett et al. analyze still pictures of spontaneous expressions collected in a 'false opinion' experiment. In both cases however, the average positive predictive rate (the percentage of actual behavior among those labeled as such by the system) in an experiment in which each behavior occurs 1% of the time would be quite low (19% for Valstar and Pantic's system, and 4% for Bartlett et al.). Even if an action unit occurs 10% of the time, the average positive predictive rate would still be much lower than the accuracy. In this scenario, between 28% (Valstar & Pantic) and 58% (Bartlett et al.) of the smiles (action unit 12, a behavior that is well represented in facial expression databases and usually among the most accurately detected) would be false alarms, i.e. other behaviors mistakenly recognized as smiles.

It should also be noted that there is a trade-off between sensitivity and false alarm rate and most systems can therefore be tweaked toward a more conservative or a more liberal decision threshold for each behavior. Published performance data are typically based on the model parameters that maximize accuracy on the learning data set. Collecting relevant movement samples and more information on actual behavior in the application situation (e.g. real product tests) is therefore a *sine qua non* to judge the practical usefulness of automatic facial expression detection.

1979; Schwartz, Fair, Salt, Mandel & Klerman, 1976). Electrodes placed on the surface of the skin can pick up electrical changes in the motor neurons innervating muscles in the area (needle electrodes can be used to increase the specificity of the measurement but given their intrusiveness they are seldom used in psychophysiological research and will not be discussed here). The intensity of contraction depends on the number of muscle fibers activated and on the rate of firing in the corresponding motor neurons. EMG therefore does not directly measure the movement itself but electrical changes associated with it (Cacioppo, Tassinary & Fridlund, 1990). Consequently, it can also record activity too small to produce visible changes detectable by observation (Cacioppo, Petty, Losch & Kim, 1986; Cohn & Ekman, 2005).

Two regions of the face have in particular been used to discriminate between positive and negative affect, corresponding to the muscles *Corrugator supercilii* and *Zygomaticus major* (while measurement areas or *loci* are generally designated by the muscle thought to dominate the signal, surface electrodes cannot strictly measure activity in a single muscle, see Fridlund & Cacioppo, 1986 for recommendations on electrode placement). *Corrugator* is a muscle drawing brows together and contributing to FACS action unit 4. *Zygomaticus major* is a muscle of the cheek, pulling lip corners up in a smile (action unit 12).

Corrugator activity has been shown to be stronger for negative stimuli in experiments with pictures of happy and angry faces, snakes and flowers, simple tones and fear conditioning (Dimberg, 1988), affective pictures (Lang et al., 1993), auditory stimuli (Bradley & Lang, 2000) and words (Larsen, Norris & Cacioppo, 2003). All these experiments have also shown an effect in the opposite direction on *Zygomaticus* activity, albeit generally smaller (Larsen et al., 2003) and not linear.

2.3.4. Use in applied research

Formal observation with the coding systems described above has been used to study facial expression in various fields of psychology (social, developmental, clinical) but not to our knowledge in applied research (be it design, music, consumer psychology, usability/HCI or media studies). A few examples of *ad hoc* observations of facial expressions in design-related research have however been published. In particular Ludden (2008) used facial expression to assess surprise in response to products breaking sensory expectations with mixed success.

While facial electromyography does require costly equipment and specialized expertise, it is still in many respects easier and cheaper than systematic coding of facial behavior and has been used in several fields of applied research.

In human-computer interaction, Hazlett (2003) found a link between *Corrugator supercilii* activity and frustration or difficulty while using a website. Mandryk and Atkins (2007) used *Zygomaticus major* and *Corrugator supercilii* EMG to compute a valence index – manually and with a fuzzy logic system combining EMG data with heart rate – and found a difference between gaming alone and with a co-located friend.

Mahlke, Minge, and Thüring (2006) found differences in *Zygomaticus major* and *Corrugator supercilii* activity between two on-screen mobile phone prototypes and weak correlations with self-report measures but *Zygomaticus* activity was higher for the most negative product, leading them to question its usefulness as a marker of positive affect. Mahlke and Thüring (2007) measured facial activity in a test of touch screen audio player prototypes, varying in ease of use and usability but found no differences in *Zygomaticus major* activity and only a weak effect of usability on *Corrugator supercilii*.

2.3.5. Interpretation issues

Coding systems – manual or automatic – or facial electromyography can provide reasonably accurate measures of visible movement or muscle activity on the face but the process underlying this behavior and its interpretation in emotion terms are far from trivial. The most influential model in this field is probably Ekman and Friesen's (1969; Ekman, 1972). In their neurocultural theory of emotion, facial expressions are part of a small set of "affect programs", one for each basic emotion. Each affect program and the associated patterns of facial movement and bodily changes are thought to be pre-wired and universal but the eliciting conditions are at least in part person- and culture-dependent. People also sometimes try to dissimulate or otherwise alter external manifestations of the affect program, especially facial expressions, following "display rules", which also are specific to a given person and culture.

Experimental support for this model would provide strong support for the validity of facial expressions measurement of emotion (see also chapter 8). While Ekman and Friesen themselves and a number of other researchers uncovered extensive data supporting it, several aspects relevant to the measurement of emotion deserve further examination.

The most hotly debated of these is the degree of universality in the facial expressions of emotion (Ekman, 1994; Izard, 1994; Russell, 1994, 1995). Both Ekman (Ekman, Sorenson & Friesen, 1969) and Izard (1971) collected data on recognition of facial expressions of basic or fundamental emotions in many different countries and cultural groups and found a broad agreement on the meaning of these

expressions. Even in isolated members of a pre-literate culture in Papua New-Guinea, Ekman and Friesen (1971) could observe above-chance recognition of anger, disgust, happiness and sadness. However, methodological artifacts (e.g. forced-choice response format) might have inflated these recognition rates and the exact meaning of these results is disputed (Russell, 1994). Still, a number of researchers obtained similar results (Ekman, 1999; Elfenbein & Ambady, 2002) and most researchers agree that facial expressions can convey some form of universally recognizable affective information (Russell, 1995).

Importantly, these results are almost exclusively based on recognition studies with *acted or imitated* expressions as stimuli. In this type of research, pictures of lay people or professional actors instructed to move their face or to play an emotion are presented to research participants and the focus of the study is on the *decoding* of these pictures by the observer. Consequently, it does not provide much information on what information is *encoded* in facial behavior, that is how frequently particular expressions occur, how often they are associated with affective processes, how often emotions occur without facial behavior, etc. Much less is known on facial expressions occurring after emotion induction or outside the lab and how much they resemble these universally recognizable basic expressions (but see Matsumoto, Keltner, Shiota, O'Sullivan & Frank, 2008, and Matsumoto & Willingham, 2006, for different studies relevant to this issue).

Another related concern is the type of emotion model that can be mapped on facial behavior and the granularity of the emotion data that can be inferred from facial measures. In recent decades, facial behavior coding systems and research on facial expression has been associated with a discrete model based on a small number of basic emotions. It was however not always so and many early studies (and some more recent, see Russell, 1995) related facial behaviors to broad dimensions of affect. Meanwhile, most facial electromyography research has also focused on valence differences, and evidence of differentiated activation for specific emotion is weak (Larsen et al., 2008). Evidence on spontaneous facial displays is also limited to broad differences between stressful and enjoyable situations (Ekman, 1999; Russell, 1994). Similarly, automatic recognition systems trained to recognize spontaneous emotions are typically based on a dimensional rather than categorical model of emotions (Pantic, 2009). It therefore appears that even if observers can recognize posed facial expressions of basic emotions, the data available only supports a dimensional model of affect for the measurement of actual emotion through facial movement.

Beside the issues of universality and specificity, more fundamental theoretical challenges against the view of facial expression implicit

in Ekman and Friesen's work have also appeared in the literature. Up to this point, the discussion was based on the assumption that facial displays simply *express* emotions, i.e. that affect directly causes muscle activity and is transparently reflected on the face. While this assumption underlies most psychological research on facial behavior and emotion and is at the core of a very fruitful research program in the psychology of emotion, it has been increasingly criticized since the 1990s (Russell & Fernández-Dols, 1997). The most distinctive alternative is Fridlund's (e.g. 1997) "behavioral ecology view", which posits that facial movement does not reflect any internal affective state but serves to communicate "social motives", i.e. intentions about the future course of the interaction (aggression, affiliation, etc.). These motives can be associated with several emotions or even with no emotion at all and the affective state of the sender plays no causal role in Fridlund's account of facial behavior. Other researchers, while retaining the notion of expression, have insisted on componential views linking facial behavior to specific facets of emotion such as appraisals (Scherer & Grandjean, 2008) or action tendencies (Frijda & Tcherkassof, 1997).

Beyond the theoretical disagreements, the most important result from this body of research is however that many other processes than emotion can influence facial movement. For example, the presence of real or imaginary observers can increase expressive behavior, independently of the strength of the emotion ("audience effects", see e.g. Fridlund, 1991). While several interpretations of these data are possible, they clearly imply that there is no more than a probabilistic connection between emotion and facial behavior (Frijda & Tcherkassof, 1997; Parkinson, 2005).

In a completely different type of research, Dimberg & Karlsson (1997) also suggested that evolutionary relevant stimuli, not valence *per se,* had an effect on *Zygomaticus major* and *Corrugator supercilii* activity. In their experiment, pictures of faces and snakes elicited stronger muscle activity in these regions than flowers or landscapes pictures, and the differences were not directly related to pleasantness and unpleasantness ratings.

Even if none of this strictly rules out any role for affect in accounts of facial behavior, these various results do in any case weaken the causal link between emotion, conceived as an inner psychological state, and movements of the face, and make any reverse inference from these facial changes to psychological processes more complex.

In fact, this conclusion is also warranted within the traditional view of facial behavior as emotion expression, even disregarding the theoretical debate about their meaning and the strength of the evidence in favor of a two-factor account. Coming back to Ekman and Friesen's model, it is easy to focus on the fact that expressions are intimately

linked with specific affect programs and to fail to appreciate that the final changes observed on the face are also the results of personal and cultural display rules. The existence of large inter-individual and inter-cultural differences in the conditions of occurrence and the meaning of facial movement is not really disputed (see e.g. Eibl-Eibesfeldt, 1997, pp. 633 sq. for a discussion of differences and universalities in eyebrow raising by an ethologist usually counted as a strong proponent of universal expressions), and the debate is really about their extent, how they should be accounted for and whether these differences are the result of another process than emotional expression *per se*.

In fact, Ekman attributed discrepancies between his results and earlier research to a failure to properly discriminate between affective behavior and other types of facial movement (Ekman et al., 1969) and suggested a number of hypotheses regarding the differences between genuine expressions of emotions and deceptive or voluntary facial displays. Unfortunately some of these hypotheses rest on limited evidence and none of them are routinely integrated in measurement strategies. For example, neither facial electromyography research with *Corrugator supercilii* and *Zygomaticus major* nor automatic recognition system trained on posed facial expression can distinguish between different types of smiles.

2.4. Measurement over time

All techniques discussed so far are typically used to obtain summary measures of affect, asking different groups of research participants to report their feelings once or comparing counts of facial expressions or mean changes in autonomic parameters over a few experimental conditions. Essentially, they probe for a respondent current affective state and can be used to collect punctual ratings of users' feelings but provide only limited information on the temporal dynamics of experience.

A number of fields have however developed instruments to measure emotional states over time and study the dynamics of affective processes, how emotions change or remain similar in relations to modifications in the environment.

These instruments can be first divided according to the time-scale considered. Researchers in developmental psychology but also in design (Karapanos, 2010) are often interested in evolutions over periods of months or years. These time scales will not be considered in this thesis, which is limited to moment-to-moment measurement during interaction sequences lasting minutes or hours.

While psychophysiological and behavioral observation techniques might seem particularly suited to this type of research because

they do not require any active involvement of research participants in the measurement process and can potentially yield enormous amounts of continuous data, studies of this kind are exceedingly rare. Psychophysiological measures for example are almost always analyzed at an aggregate level, comparing means or peaks between different conditions (e.g. tasks, pictures, films) without much attention to the dynamics of the process (for an exception see Ravaja et al., 2008). Numerous repetitions (e.g. several pictures of the same valence) are often used to compensate the noisiness of the measurement. The review will consequently focus on self-report instruments developed specifically for this purpose.

Aaker, Stayman, and Hagerty (1986) introduced such a procedure, called the "warmth monitor", in advertising research. Stayman and Aaker (1993) collected data supporting test-retest reliability, convergence with skin conductance and post-advertisement adjective ratings and establishing that "warmth" was not simply "liking" (but see Vanden Abeele & MacLachlan, 1994, for a criticism of these results). Studies using these techniques continue in advertisement research, for example to investigate the effect of experience on the probability to stop viewing (Woltman-Elpers, Wedel & Pieters, 2003). Biocca, David, and West (1994) discuss several studies of "communicative messages" with a similar instrument, the continuous response measurement (in practice a small rating dial). They use it to collect both affective (mood) and cognitive (evaluations, opinions) reports from participants watching a message.

Gottman & Levenson (1985) used a big rating dial (rotating on 180°) to collect self-report of affect from spouses involved in low-conflict and high-conflict interactions (see Ruef & Levenson, 2007, for details about the device and procedure and a discussion of analysis strategy).

In music education and music perception research, continuous rating of various perceptual dimensions has also become very popular. The most widely used tool for this kind of research is probably the Continuous Response Digital Interface (CRDI); according to its developers it has been used in more than 70 studies (Geringer, Madsen & Gregory, 2004). Rather than a specific instrument, the CRDI is in fact a series of devices that can be combined with different instructions to define a family of continuous measurements. The first CRDI was a large dial that could be rotated over 256 degrees. Recent versions took the form of a box with a lever than can be moved back and forth (direction can be changed by placing the box differently). In most studies, the meaning of the scale is defined through the instructions and by placing various overlays on the CRDI.

This approach makes comparing reliability or validity across study impossible and raises questions regarding the discriminant validity of

the CRDI. For example, Lychner (1998) found that participants asked to report their experience of music in terms of "aesthetic response" or "felt emotional response" provided very similar ratings, while "tension" was clearly different from the rest of the data. Despite being ostensibly different things, "aesthetic response" and "felt emotional response" therefore seem to be understood similarly by research participants.

Schubert (1999) developed a software-based self-report instrument called "two-dimensional emotion-space" (or 2DES) to address concerns with the specificity of one-dimensional tools and presented several careful validation studies with music excerpts. Participants have to move the mouse cursor in a valence/arousal space anchored by schematic faces (with the shape of the mouth representing valence and the size of the eyes and mouths representing arousal). EMuJoy (Nagel, Kopiez, Grewe & Altenmüller, 2007) and Feeltrace (Cowie et al., 2000), or the AffectButton (Broekens, Pronker & Neuteboom, 2010) are very similar tools with a more up-to-date user interface. Both can be downloaded on the web.

A few results from this literature could have considerable import for research on the dynamics of experience if they could be replicated or extended in product use situations. One of these pertains to the link between moment-to-moment ratings and overall evaluation of an experience. In two separate studies of this question, Brittin & Duke (1997) and Duke & Colprit (2001) found that summative ratings collected after the fact and mean continuous ratings of particular musical excerpts were consistent across participants but differed systematically from each other. These findings suggest that continuous self-report does indeed provide information that is not equivalent to overall ratings. This is also coherent with research on the role of peak and end experience on the formation of global impressions (Fredrickson & Kahneman, 1993). However, working with recruitment messages, Reeve, Highhouse & Brooks (2006) collected data providing more support to an averaging model than to the peak-end rule.

2.5. Conclusion

This literature review identified many measures of emotion. Among them, self-report of conscious feelings is certainly the most common and versatile technique. Self-report questionnaires based on different models of affect have been used in design-related research. Many of these questionnaires were however initially developed as measures of moods and only measure diffuse feelings of pleasantness and unpleasantness rather than specific responses to an object. Additionally, measures derived from the psychological or clinical literature have also

been criticized for their excessive focus on negative affect. Several questionnaires have been developed to address these limitations, most notably PrEmo.

While some of these questionnaires would seem relevant to the measurement of emotional experience in person-product interaction, measurement-oriented publications are often limited to research on product appearance or surveys about attitudes toward recently bought products. Chapter 3 addresses this deficiency by presenting two studies in which emotion was measured immediately after interacting with a product.

Self-report was also used to collect moment-to-moment ratings of feelings in several fields but the techniques described in the literature require constant interaction between the research participant and the data collection apparatus. Chapter 4 presents a new approach combining these moment-to-moment self-report procedures with video to be able to study minute changes in feelings during interaction with products.

Emotion measures based on other components than conscious feelings have also been extensively discussed in several applied fields. Since they can continuously record minute changes with a high sensitivity, these techniques would seem more suited than self-report for moment-to-moment assessment but actual studies of the dynamics of emotion using physiological or behavioral recording are in fact very rare, possibly because the complexity of the apparatus and data analysis and because the lack of reliability of these measures makes averaging over multiple trials almost unavoidable. While the promise to index unconscious processes and to eschew reliance on participants (self-) conscious reports is enticing, empirical evidence on the usefulness of these techniques remains limited and they suffer from a number of interpretation difficulties. Chapter 8 provides an extensive discussion of these issues.

Finally, the review also identified several findings on the formation of overall impressions based on ongoing experience that could have important consequences for interaction design if they could be extended to user experience with products. Chapter 5 shows how the techniques developed in this thesis can be combined to tackle this question and presents a first attempt at generalizing these effects to design-oriented research.

3. Questionnaire Assessment of Emotional Experience

Despite the large of number of tools, approaches, and instruments developed to measure emotions and the amount of interest for user experience and emotions elicited by products, there are actually very few studies looking at the empirical characteristics of these measurement procedures within the context of *interactive* product design. The studies that do exist and are documented in the literature often focus on product appearance or perception (participants are shown a product and asked to provide ratings or otherwise react to it without actually using it for its intended purpose) or on general satisfaction (participants are asked, perhaps in a survey, to rate some products they have used in the past).

The present chapter discusses two studies in which the experimenter provides a product and participants are asked to actually use it. The main goal of these studies was to test the sensitivity of several emotion questionnaires to this manipulation but the emotion data will also be related to other aspects of user experience and the results will be used as a reference when discussing the dynamics of experience (chapter 5) and the reliability and validity of emotion measures (chapter 7 and 8).

3.1. Experiment 1: Coffee machine/alarm clock[1]

The first of these two experiments compared self-reported ratings of emotional experience after using two products (a coffee machine and an alarm clock) with two different questionnaires. These two questionnaires were selected because they cover many different positive emotions and come from leading research groups in design and emotion psychology (see also chapter 2).

The first of these questionnaires was Desmet's (2004) PrEmo. It was developed to measure people's response to product appearance

1 Most of the material in this section was published in the proceedings of *Design and emotion 2008*. This paper was nominated for a best paper award at the conference. I am thankful to David Güiza Caicedo and Marleen van Beuzekom for their help in organizing the study and collecting the data.

and consists of 10 purely non-verbal single-item scales. Each of these items consists of an animated cartoon representing a particular emotion using facial expression, body movement and sound.

The second questionnaire, the Geneva Emotion Wheel (GEW) is a more traditional self-report questionnaire using words (emotion names) as item labels. It was not developed specifically for design research but, unlike many emotion measures from psychology, covers a large number of positive and negative affect states represented by single-item scales. Self-report instruments based on discrete emotions tend to be *ad hoc* questionnaires and adjective lists, harming the comparability between studies and the accumulation of knowledge in this field. The GEW was developed to improve on this situation, by the design of a questionnaire going beyond the valence-arousal space but organizing verbal labels in a systematic fashion that would make the tool easier to use, and more reliable across studies (Scherer, 2005).

The current version of the Geneva Emotion Wheel consists in a set of 20 emotion families, selected among those most studied in the field or considered as "basic emotions". These emotion families are organized in a circle, but instead of grouping them according to the traditional valence and arousal dimensions, their position is determined by fundamental appraisal dimensions. The vertical axis represents the power/control appraisal and the horizontal axis the pleasantness appraisal. The Geneva Emotion Research Group provides English, French and German-language versions of the GEW.

An initial Dutch translation was prepared by Pieter Desmet and subsequently revised with the help of another Dutch-speaking emotion researcher (Johnny Fontaine, University of Leuven) and one of the authors of the original questionnaire (Klaus Scherer, University of Geneva). As in the English-language version of the GEW, items include both nouns (e.g. "irritation", "schaamte") and verbs ("feeling disburdened", "genieten"). Table 3.1 lists all items in Dutch and English.

Table 3.1. *Translation of the emotion families of the Geneva Emotion Wheel.*

High control/Low pleasantness		High control/High pleasantness	
English	Dutch	English	Dutch
Irritation	Irritatie	Involvement	Betrokkenheid
Anger	Boosheid	Interest	Interesse
Contempt	Minachting	Amusement	Amusement
Scorn	Bitterheid	Laughter	Lachen
Disgust	Walging	Pride	Trots
Repulsion	Weerzin	Elation	Verrukking
Envy	Afgunst	Happiness	Geluk
Jealousy	Jalousie	Joy	Blijheid
Disappointment	Teleurstelling	Enjoyment	Genieten
Regret	Spijt	Pleasure	Plezier
Low control/Low pleasantness		**Low control/High pleasantness**	
English	Dutch	English	Dutch
Guilt	Schuldbewust	Tenderness	Genegenheid
Remorse	Berouw	Feeling love	Liefde voelen
Embarrassment	Gegeneerdheid	Wonderment	Verwondering
Shame	Schaamte	Feeling awe	Ontzag voelen
Worry	Verontrusting	Feeling	Bevrijd voelen
Fear	Angst	disburdened	
		Relief	Opluchting
Sadness	Bedroefdheid	Astonishment	Verbazing
Despair	Vertwijfeling	Surprise	Varrassing
Pity	Medeleven	Longing	Verlangen
Compassion	Medeogen	Nostalgia	Nostalgie

3.1.1. Material and methods

The participants (N = 40) were students in Industrial Design at Delft University of Technology, all of them native Dutch speakers. They were asked to use two products and to report about their experience with both questionnaires after using each product. The products were chosen for their potential to elicit different emotions. One of them was a Phillips/Alessi designer coffee machine, expected to be pleasant to use because of its function and its overall design. The other one a rather complex alarm clock, providing for a rather frustrating experience. As appraisal theories underline the importance of goals and concerns in affective responses (Desmet & Hekkert, 2002), participants were asked to carry out a task with each product (brew coffee and set up an alarm).

Figure 3.1. *Screenshot of the Dutch version of the Geneva Emotion Wheel as it was presented to the participants. In this example, five emotion families are selected with various levels of intensity. Instructions read "Use the above scales to describe your feeling toward the product (more than one choice is possible). Emotion intensity scale: low … high"*

After using each product, the participants were asked to report their feelings using two questionnaires: the Dutch translation of the GEW presented above and PrEmo (figures 3.1 and 3.2). Both questionnaires were administered on-screen using custom-made software developed with Adobe Flash.

In keeping with the original paper-and-pencil response sheet, the different items of the GEW were displayed all at once in a circular

format (Scherer, 2005; Tran, 2004). This wheel or circle is not based on the traditional valence/arousal circumplex (Russell, 1980), but on two of Scherer's "stimulus evaluation checks" (Scherer, 1984, as cited in Scherer, 2005). The vertical axis thus corresponds to the "control" dimension whereas the horizontal axis reflects the level of "pleasantness" of each emotion. Participants could select any number of emotions and indicate the level to which they experience each of these emotions on a five-point scale going from the inside toward the outside of the circle. It was therefore also possible to select only a few items in the wheel and let the other untouched (implicit "not at all" position).

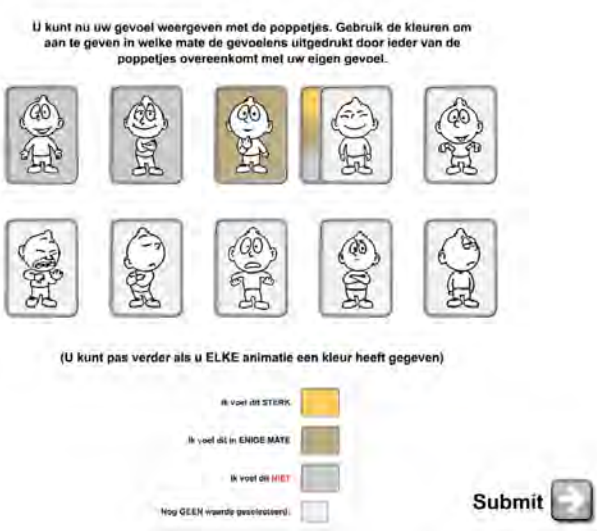


Figure 3.2. *Screenshot of PrEmo as it was presented to the participants. Instructions read: "You can now render your feeling with the animated characters. Use the colors to indicate to which extent the feelings portrayed by each character corresponds to your own feeling. (You can only proceed further after giving a color to each animation)"*

The version of PrEmo used in this study (figure 3.2) is a ten-emotion version similar to the one used in Desmet, Porcelijn & van Dijk (2007). The emotions included are positive surprise, satisfaction, fascination, amusement, desire, disgust, contempt, negative surprise, dissatisfaction and boredom. These labels correspond to the researcher's description of the emotions portrayed and were also validated in a study involving Japanese, US, Finnish and Dutch participants (Desmet, 2002) but they are not presented to the participants, who have to rate their experience based solely on the animations, without verbal description of the emotions. For each of the ten animations, participants had to indicate how closely it matched their feelings with a three-points scale ("Ik voel dit STERK" – I am feeling this strongly, "Ik voel dit in ENIGE

MATE" – I am feeling this somewhat, "Ik voel dit NIET" – I am not feeling this).

To reduce spillover and learning effects, the order of products and questionnaires was counterbalanced. Half of the participants were asked to use the coffee machine first, while another half had to set up the alarm clock first. In each group, half of the participants used the GEW first and the other half began to report their feelings with PrEmo (table 3.2).

Table 3.2. *Overview of experimental design*

First product used	First questionnaire	N
Coffee machine	PrEmo	10
	GEW	10
Alarm clock	PrEmo	10
	GEW	10

3.1.2. Results

Unlike many mood questionnaires discussed in chapter 2, the two questionnaires used in this experiment were not designed to assess two or three dimensions but as measures of discrete emotions. Each animation or pair of emotion words can thus be understood as a single-item scale. Still, as explained in section 2.1.3, emotion data can be interpreted through a hierarchical structure going from discrete emotions to an overarching bipolar valence dimension. Ratings of discrete emotions therefore should not be expected to be totally independent and even questionnaires that have not been devised factor-analytically to measure this underlying valence dimension might be used to derive a pleasantness index. The data from PrEmo and the GEW will accordingly be analyzed at all three levels of Tellegen, Watson & Clark (1999) hierarchical structure of affect.

The first level of the hierarchy is formed by categorical or discrete emotions like happiness, anger/irritation, and disgust. PrEmo was developed as a measure of 10 to 14 of these discrete emotions, thought to be the most relevant for design stimuli. The GEW includes a larger set of 20 emotions selected to comprehensively cover the emotions most often discussed in the literature. At this level of analysis, it is difficult to assess the convergence between the two instruments, as there are 435 possible correlations between the 30 items of both questionnaires combined. Such a large correlation table is unwieldy to report and interpret, certainly with such a limited sample size.

It is however possible to examine individual items scores emotion-by-emotion to find out if the two products elicited different rating. As shown in table 3.5, many of these differences are significant, with

the strongest ones for "enjoyment" (GEW) and "irritation" (GEW). The emotions showing no significant difference between products are "amusement" (PrEmo and GEW), "negative surprise" (PrEmo), "pride" (GEW), "guilt" (GEW), "regret" (GEW), "relief" (GEW), "astonishment" (GEW), "longing" (GEW), "pity" (GEW), "worry" (GEW) and "envy" (GEW)[2].

2 The magnitude of the differences between product on individual GEW and PrEmo items are not directly comparable because of the dissimilar response formats.

Table 3.3. *Item-by-item comparisons between coffee machine and alarm clock.*

Emotion	Alarm clock	Coffee maker	Difference		
	Mean (SD)	Mean (SD)	Raw diff.	P-value (adjusted)	Correlat.
PrEmo					
Positive surprise	0.7 (0.7)	1.3 (0.8)	- 0.55	.00 (.05)	-.05
Satisfaction	0.7 (0.6)	1.3 (0.7)	- 0.58	.00 (.00)	.26
Fascination	0.4 (0.6)	1.0 (0.7)	- 0.58	.00 (.00)	.47
Amusement	0.4 (0.7)	0.5 (0.6)	- 0.08	.62 (1)	-.10
Desire	0.5 (0.7)	0.7 (0.7)	- 0.23	.05 (.77)	.51
Disgust	1.1 (0.8)	0.3 (0.5)	0.80	.00 (.00)	-.03
Contempt	0.7 (0.7)	0.3 (0.5)	0.40	.01 (.11)	-.03
Negative surprise	0.7 (0.8)	0.4 (0.7)	0.10	.49 (1)	.20
Dissatisfaction	1.0 (0.9)	0.3 (0.6)	0.68	.00 (.00)	.26
Boredom	0.7 (0.6)	0.4 (0.6)	0.25	.05 (.77)	.18
GEW					
Involvement	1.5 (1.7)	2.1 (1.7)	- 0.63	.09 (1)	.12
Amusement	0.7 (1.2)	1.2 (1.5)	- 0.53	.07 (.86)	.17
Pride	1.2 (1.7)	1.5 (1.6)	- 0.35	.26 (1)	.33
Happiness	0.5 (1.0)	1.2 (1.5)	- 0.70	.01 (.14)	.24
Enjoyment	0.6 (1.3)	1.8 (1.7)	- 1.28	.00 (.00)	.24
Tenderness	0.1 (0.4)	0.3 (0.9)	- 0.23	.11 (1)	.41
Wonderment	0.4 (1.0)	1.4 (1.6)	- 0.95	.00 (.01)	.43
Relief	0.7 (1.4)	0.6 (1.2)	0.10	.71 (1)	.17
Astonishment	2.0 (1.7)	2.4 (1.7)	- 0.38	.25 (1)	.27
Longing	0.4 (1.0)	0.5 (1.1)	- 0.15	.39 (1)	.45
Pity	0.2 (0.7)	0.2 (0.5)	0.03	.83 (1)	.33
Sadness	0.9 (1.3)	0.1 (0.2)	0.80	.00 (.01)	.16
Worry	0.4 (0.9)	0.7 (1.3)	- 0.23	.32 (1)	.16
Shame	1.0 (1.5)	0.2 (0.6)	0.78	.00 (.08)	.03
Guilt	0.1 (0.5)	0.1 (0.2)	0.08	.08 (1)	.94
Regret	1.0 (1.4)	0.5 (1.2)	0.48	.06 (.77)	.31
Envy	0.2 (0.8)	0.0 (0.2)	0.18	.18 (1)	.16
Disgust	1.2 (1.5)	0.3 (.8)	0.90	.00 (.05)	-.06
Scorn	1.3 (1.5)	0.5 (1.0)	0.83	.00 (.08)	.06
Irritation	2.9 (1.7)	0.3 (0.8)	2.53	.00 (.00)	.07

(Unadjusted) p-values correspond to paired T-tests with 39 degrees of freedom, testing whether product mean scores on each item differ. Adjusted p-values are computed with Holm's procedure to control the family-wise error rate for all tests in this table (Shaffer, 1995; Wright, 1992). The last column represents the correlation between ratings for the coffee maker and the alarm clock and can be used for effect size and power calculations.

The second level in the hierarchy is probably more appropriate to assess the level of convergence between both questionnaires. In Tellegen, Watson & Clark (1999) model, the intermediate level is dominated by two distinct unipolar dimensions: positive and negative activation The usual way to derive positive and negative affect scores from discrete emotion ratings is to use some form of factor analysis. In this study however, the modest sample size and the characteristic of the data matrix suggest that such a strategy might not be appropriate[3].

A visual inspection of the overall correlation matrix does however suggest that there are some meaningful associations between emotions of the same valence[4]. For PrEmo ratings in particular, the strongest correlations are observed between different positive emotions or between different negative emotions. Moderate negative correlations are also apparent between emotions of opposite valence. It was therefore decided to group PrEmo emotion in two 5-item parcels, defined a priori by valence rather than through factor analysis. This bidimensional structure also agrees well to theoretical expectations derived from influential models of affect (see chapter 2, section 2.1.3). For the GEW, the structure is rather unclear and the emotions have been grouped in four quadrants, following Tran (2004). The four groups represent achievement emotions (high control, high pleasantness emotions like enjoyment and pride), approach emotions (low control, high pleasantness emotions like interest and surprise), resignation emotions (low control, low pleasantness emotions like sadness and shame), and antagonistic emotions (high control, low pleasantness emotions like disgust and anger). Table 3.3 and 3.4 show the resulting correlation matrices[5].

3 The Kaiser-Meyer-Olkin measure of sampling adequacy is 0.475 for the alarm clock data and 0.352 for the coffee machine ratings, well under the acceptable limit of 0.5 or 0.6 and the matrix determinants are also dangerously small (both smaller than 10-11).

4 Ratings for each product were analyzed separately to ensure that each observation is independent (i.e. each participants contributes a single pair of observations to each correlation coefficient and all observations used in the analysis refer to the same product, which would not be the same if the data were pooled) and precludes a range of interpretation problems explained in more details in chapter 7. Unfortunately, it also means that the correlations reflect the variation between participants (in response to one product or in general) but not necessarily within-participant differences between products.

5 All correlation coefficients are Kendall's τ coefficients, as it is recommended as replacement for Pearson's r for non-normal data and small samples with a high number of ties.

Table 3.4. *Correlations (Kendall's τ) between item parcels for the alarm clock.*

	1	2	3	4	5	6
1. PrEmo positive emotions	1					
2. PrEmo negative emotions	-.36	1				
3. GEW high control/pleasant	.52	-.35	1			
4. GEW low control/pleasant	.41	-.23	.32	1		
5. GEW high control/unpleasant	-.36	.59	-.37	-.11	1	
6. GEW low control/unpleasant	.03	.14	.09	.30	.22	1

For both products, there are relatively strong associations between positive emotions in PrEmo and the GEW (both low and high control) and between negative PrEmo emotions GEW unpleasant emotions (except low control emotions for the alarm clock). These associations support the distinction between two basic types of emotions, pleasant and unpleasant.

PrEmo positive emotions also show a moderate negative correlation with PrEmo negative emotions and with high control/low pleasantness emotions in the GEW. These negative correlations are consistent with the idea of a higher-order bipolar valence dimension. These patterns are very similar in both products.

Finally, GEW emotions with the same level of control but opposite valence also show a modicum of association. However, correlations between GEW low control/unpleasant emotions and all other groups of emotions tend to be lower. This lack of association with other variables is likely due to the fact that participants rarely used these items, thus reducing score variance and attenuating any possible correlation.

Table 3.5. *Correlations (Kendall's τ) between item parcels for the coffee machine.*

	1	2	3	4	5	6
1. PrEmo positive emotions	1					
2. PrEmo negative emotions	-.30	1				
3. GEW high control/pleasant	.48	-.36	1			
4. GEW low control/pleasant	.46	-.12	.28	1		
5. GEW high control/unpleasant	-.29	.49	-.42	-.15	1	
6. GEW low control/unpleasant	.03	.36	.06	.20	.29	1

As noted before, the differences in emotion ratings between the coffee machine and the alarm clock provide a test of the relevance of these measures for design-related research. If the tools compared here are able to measure product emotions, they should discriminate between the two products. This can also be assessed at the highest level of the hierarchy to confirm that the valence of participants' emotional experience corresponded to the hypotheses about each product.

For the last part of the analysis, PrEmo was therefore treated as a single valence scale and an overall pleasure-displeasure score was computed by adding the individual scores on each of the 10 PrEmo items. "Not at all" was coded 0, "a little" 1 and "strongly" 2. Ratings for negative emotions (dissatisfaction, disgust, etc.) were inverted so that a higher scale score would mean more positive and less negative emotions (theoretically, the minimum score is 0 and the maximum is 20). There is a significant difference in overall emotional experience between the coffee maker and the alarm clock, $t(39) = 5.78$, $p < .001$, 95% CI for the difference: [2.75, 5.70]. The average PrEmo score for the alarm clock ($M = 8.8$, $SD = 3.7$) is markedly smaller than the average for the coffee maker ($M = 13$, $SD = 3.2$).

Since the experiment used a within-subject design, a follow-up analysis was conducted to alleviate concerns about order effects and obtain an unbiased estimate of the main effect. The ratings of the first trial by each participant (i.e. the first product they saw during the session) were analyzed separately with an independent sample t-test (Maxwell & Delaney, 1990). This analysis "throws out" half of the data and would consequently be expected to be less powerful but completely rules out any type of transfer or interaction between the conditions, as participants had only seen a single product before providing these ratings. It is in effect treating the first set of ratings as a between-subject experiment, as if participants did not use and evaluate a second product afterwards. Even in this case, the difference in PrEmo ratings between the alarm clock ($M = 7.7$, $SD = 3.4$) and the coffee machine ($M = 13.5$, $SD = 3.15$) is significantly different from 0, $t(37.75) = 5.53$, $p < .001$, 95% CI for the difference: [3.64, 7.86]. Mean scores for each product when tested first or second are represented on figure 3.3.

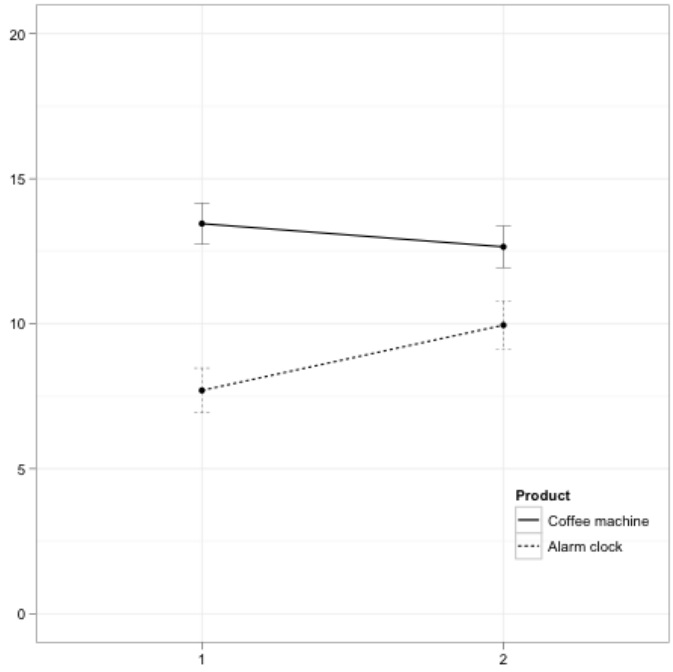


Figure 3.3. *Mean PrEmo scores for the coffee machine and alarm clock when tested first (left) and second (right).*

3.1.3. Discussion

Overall, these results show a great deal of correspondence between both instruments. The score differences between products also establish the sensitivity of both questionnaires to user experience differences. Despite the fact that both PrEmo and the GEW were designed to assess discrete emotions rather than underlying dimensions, these dimensions (and in particular pleasantness) are clearly apparent in the data. Because of the limited number of products tested, it is more difficult to reach conclusions on individual emotions but a number of observations are still possible.

Several GEW items were rarely used by participants and seemed less relevant to the product-use situation studied in this experiment. The lack of variance in scores for these emotions was in turn reflected in lower correlations with other emotions and a lack of differences between products. This was in particular the case of the low control/low pleasantness emotions guilt, embarrassment, worry, and pity (called "resignation emotions" by Tran, 2004) but also of a few other emotions such as longing, tenderness, and envy. Several PrEmo

emotions (boredom, amusement, desire, negative surprise) also exhibited little or no differences between products.

The lack of noticeable differences between products for some emotions might be explained by the specific choice of stimuli for this experiment. For example, surprise-related emotions such as negative surprise (PrEmo) or astonishment (GEW) have been shown to be elicited by products (Ludden, 2008) but did not clearly differentiate the two products in this study. Several other (pity, envy, pride, tenderness) are social emotions, typically associated with interpersonal relationships. While products can elicit this type of emotions (Desmet, & Hekkert, 2002), they were apparently less directly relevant to the products at hand. Interestingly, most of these emotions are not included in PrEmo, a tool developed specifically to measure design-related emotions.

3.2. Experiment 2: Personal navigation devices

The second experiment presented in this chapter compared users' experience with personal navigation devices for cars. All products used in the study therefore belonged to the same category, as is typically the case in tests and evaluations performed during product development. By contrast, the coffee machine and alarm clock used in the previous experiment could be expected to elicit very different experiences but the magnitude of this difference would not be representative of the kind of effects practitioners might encounter when comparing different design alternatives for the same product. Extending results to within-category differences and establishing sensitivity to the differences between relatively similar designs is therefore necessary before making claims about the usefulness of a measurement instrument in product development.

The study also took place within a larger research project[6] aiming at developing measures of several aspects of user experience, including meaning, aesthetics and emotions (Desmet & Hekkert, 2007). A pre-study led to the selection of a number of adjectives related to these experiences for a self-report questionnaire covering all three facets. Another study tested the structure of ratings of different personal navigation devices with this questionnaire. The devices were presented to 28 consumers in a lab using photographs and videos (Desmet & Schifferstein, 2010)[7].

6 This project was set up in partnership with Renault.

7 While all interpretations presented here are mine, I was not involved in the development of the questionnaire and the first study, which

The last study of the project, presented in the remainder of this chapter, aimed at assessing the same aspects of user experience after actually using the navigation devices as opposed to simply manipulating it and watching a video of someone else driving with it[8]. Beside the questionnaire developed in the course of this research project, it also included several other measures targeting various aspects of user experience, including hedonic quality and perceived usability. Moment-to-moment ratings with the self-confrontation procedure were also collected, but these data will be described in chapter 5, section 5.2.

3.2.1. Material and methods

The products used were three personal navigation devices representing a range of manufacturers and map designs: Mio Moov 580, Blaupunkt TravelPilot 500, and TomTom XL (figure 3.4). All three devices were used in previous research and shown to differ in perceived usability and user experience. Each of them has a distinctive look and feel: the TomTom XL has a straightforward no-frills graphic design with a flat pseudo 3D map, the Mio Moov uses a 3D view of the surroundings and the Blaupunkt Travel pilot is an augmented reality device, showing direction instructions superimposed on a live image from a camera placed on the back of the device (i.e. facing the front of the vehicle, when attached on the windshield).

Figure 3.4. *Stimuli used in experiment 2, from left to right:TomTom, Mio and Blaupunkt navigation devices.*

Forty participants (31 men and 9 women, aged between 20 and 55, $M = 26$, $SD = 7$ years) were recruited through posters, leaflets placed on cars parked on the campus and word of mouth. Precondition for participation was to hold a driver's license and have access to a car.

After welcoming the participants and explaining the purpose of the experiment, a camera was installed on the back seat (see chapter 5 for

was planned and conducted by Pieter Desmet and Rick Schifferstein (see Desmet & Schifferestein, 2010). Its results will therefore not be reported in detail.

8 I am very thankful to Lara van der Veen for her great help during the preparation and data collection for this study.

details on this part of the data). All participants were asked to follow the same route to a little known part of town with their own car using one of the three personal navigation devices, preprogrammed by the moderator. Once they reached the goal, participants were asked to enter a new address in the device using a detailed instruction sheet and to return to the university. A parking spot was reserved to ensure easy access to the lab, where the different questionnaires were administered before proceeding to the video-supported moment-to-moment self-report (for more detail on this part of the experiment see chapters 4 and chapter 5, section 5.2). Brief mood self-ratings using the self-assessment manikin (Bradley & Lang, 1994) were also collected in the car at four points during the drive: before starting, right after stopping at the first destination, after entering the second destination, and finally after parking the car at university. The whole drive took between 20 and 35 min (with an average of 25 min).

After returning to the lab, participants filled in four questionnaires about their experience: the Simple Usability Scale (Brooke, 1996), AttrakDiff (Hassenzahl, 2004; Hassenzahl, Burmester & Koller, 2003), PrEmo (Desmet, 2002) and the adjective-rating questionnaire developed in the course of the research project. Both the Simple Usability Scale and AttrakDiff were translated into Dutch based on the original English-language version. The translations were subsequently revised based on a back-translation and, in the case of AttrakDiff, on comparison with the German-language version[9].

The Simple Usability Scale is a Likert scale designed to assess the level of usability perceived by users of a product (i.e. the subjective or "satisfaction" component of usability, as defined by ISO-9241). It was slightly modified to adopt a response format closer to the other questionnaires, namely 7-point ratings from "disagree" ("*oneens*") to "agree" ("*eens*")[10].

The version of AttracDiff used in this study is a 28-item semantic differential questionnaire. It consists of pairs of adjectives like "human – technical" ("*menselijk – technisch*") or "simple – complicated" ("*eenvoudig – ingewikkeld*") and comprises four scales: pragmatic quality, stimulation, identification and a general attractiveness scale. Stimulation and identification are two types of hedonic attributes. The hedonic quality-stimulation scale is related to the experience of novelty and challenge while the hedonic quality-identification scale reflects the link between a product and different values or self-images.

The emotion questionnaire used in this study is identical to the one used in Desmet & Schifferstein (2010). It is based on PrEmo but uses

9 I am grateful to Jeroen Arendsen for making the initial translated version available to me.

10 The scaling factor used by Brooke (1996) was also adjusted to keep the final summative score in the 0-100 range.

a slightly different set of emotions and a different format. To integrate it with the other questionnaires in a pen-and-paper procedure, the items were reduced to a still picture of each expression together with a word describing the corresponding emotion, as opposed to the purely non-verbal animations used in other PrEmo studies. The emotions included were contempt ("*minachting*"), dissatisfaction ("*ontevreden*"), unpleasant surprise ("*onaangenaam verrast*"), rejection or disgust ("*afkeer*"), boredom ("*verveling*"), sad ("*droevig*"), admiration ("*bewondering*"), satisfaction ("*tevreden*"), pleasant surprise ("*aangenaam verrast*"), attraction or desire ("*aantrekking*"), fascination ("*fascinatie*"), and joy ("*blij*"). In keeping with earlier studies, the questionnaire uses a 3-point response format, "I don't feel this" ("*dit voel ik niet*"), "I am feeling this a little" ("*dit voel ik een beetje*"), and "I am feeling this strongly" ("*dit voel ik sterk*").

Finally, the meaning questionnaire developed in the earlier phase of the project uses a 24-item adjective-rating format (Desmet & Schifferstein, 2010). The instructions asked how well each word described the product with a 7-point response format going from "not" ("*niet*") to "very" ("*wel*"). The items and some possible English translations are listed in table 3.6.

Table 3.6. *Items and translation for "meaning" questionnaire.*

Item	Translation
Behulpzaam	Helpful, attentive
Handig	Handy, convenient, clever
Duidelijk	Clear
Slim	Smart, clever
Gebalanceerd	Balanced
Betrouwbaar	Reliable
Stimulerend	Stimulating
Interessant	Interesting
Zakelijk	Business-like, professional
Stoer	Tough, sturdy
Stijlvol	Stylish
Authentiek	Authentic
Eigenzinnig	Headstrong, stubborn
Intimiderend	Intimidating
Overdadig	Abundant, excessive
*Opvall*end	Striking, distinctive
Speels	Playful
Onrustig	Restless
Ouderwets	Old-fashioned
Goedkoop	Cheap
Abstract	Abstract

3.2.2. Results

Before comparing the different products and scales included in the study, a component analysis, reported in appendix B, was conducted to investigate the structure of the adjective questionnaire. Based on the results of this analysis, two summative scales were devised. The scores for the first scale, called helpfulness, were computed by adding item ratings for "helpful", "handy", "stimulating", "smart", "clear", "reliable", "balanced", and "abundant". The scores for the second scale, called distinctiveness, were obtained by adding the ratings for "cheap", "distinctive", "playful", and "old-fashioned". Scores for all scales (including AttrakDiff) were rescaled to fall between 0 and 100 for convenience. The average scores per product on each scale will be compared using separate one-way ANOVAs[11].

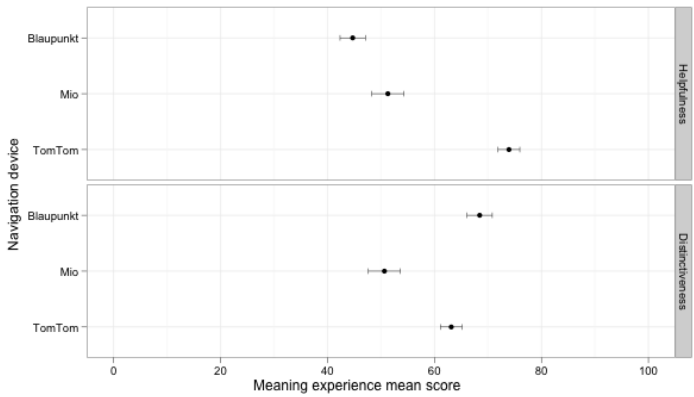


Figure 3.5. *Mean "helpfulness" and "distinctiveness" ratings for each navigation device (error bars: standard error of the mean)*[12].

As shown in figure 3.5, the mean helpfulness score for the TomTom device is the highest (M = 74, SD = 15), followed by the Mio (M = 51, SD = 22) and the Blaupunkt (M = 45, SD = 19). Together, these differences are significant; $F(2, 37) = 8.48$, $p < .001$. The order of the mean distinctiveness scores for the three devices is different; this time the Blaupunkt navigation device has the highest score (M = 68, SD = 15) together with the TomTom (M = 63, SD = 13) followed by the Mio (M = 51, SD = 19). These difference is also significant, $F(2, 37) = 4.50$, $p = .018$.

11 Performed with R *aov* function.

12 All statistical graphs in this thesis have been prepared with GGplot2 (Wickham, 2009).

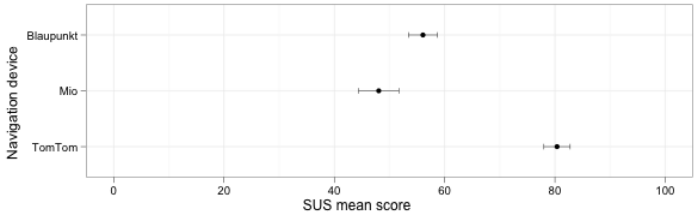


Figure 3.6. *Mean perceived usability for each navigation device (error bars: SEM).*

Usability ratings with the Simple Usability Scale[13] (figure 3.6) also reveal a difference between TomTom (M = 80, SD =15), Blaupunkt (M = 56, SD = 16), and Mio (M = 48, SD = 23), $F(2,36) = 11$, $p < .001$.

13 The data from one participant (using the Mio Moov 580) were not included in the analysis because of a missing rating for the item "*Ik vond dat er teveel tegenstrijdigheden om dit navigatiesysteem zaten*" ("I thought there was too much inconsistency in this system").

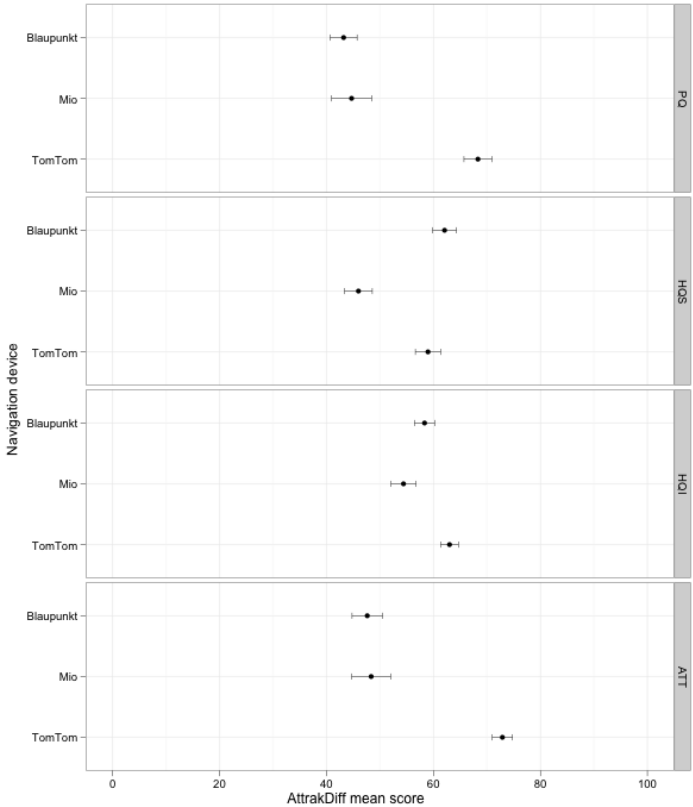


Figure 3.7. *Mean scores for each navigation device on AttrakDiff's user experience scales (error bars: SEM). PQ = Pragmatic quality, HQS = Hedonic quality – Stimulation, HQI = Hedonic quality – Identification, ATT = Attractiveness.*

Figure 3.7 presents the mean scores of each personal navigation device on AttrakDiff's various user experience scales. The TomTom navigation device has the highest mean score on AttrakDiff's Pragmatic Quality scale ($M = 68$, $SD = 17$). For the same scale, there is virtually no difference between the Mio ($M = 45$, $SD = 24$) and the Blaupunkt ($M = 43$, $SD = 16$). An omnibus test of the differences between all three devices is significant, $F(2, 37) = 7.29$, $p = .002$. There are also some significant differences in mean Hedonic Quality – Stimulation scores, $F(2, 37) = 4.23$, $p = .022$. Highest scoring products are the Blaupunkt ($M = 62$, $SD = 14$) and the TomTom ($M = 59$, $SD = 15$) with the Mio scoring lowest ($M = 46$, $SD = 16$). Hedonic Quality – Identification scores are not very different from one navigation device to the other (Mio: $M = 54$, $SD = 15$; Blaupunkt: $M = 58$, $SD = 12$; TomTom: $M = 63$, $SD = 10$) and all around the middle of the scale, $F(2, 37) = 1.55$, $p = .23$. Scores for the attractiveness scale are very

similar to the Pragmatic Quality scores, with the TomTom first (M = 73, SD = 12) followed by the Mio (M = 48, SD = 23) and Blaupunkt (M = 48, SD = 18), omnibus test in ANOVA: $F(2,37) = 8.24$, $p = .001$.

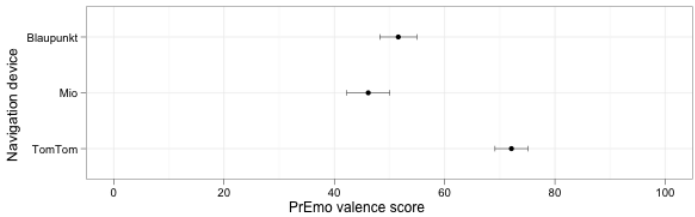


Figure 3.8. *Mean PrEmo emotion/valence score for each personal navigation device (error bars: SEM).*

Finally, a simple emotion (valence) score was computed by adding ratings for all PrEmo items, after inverting the scores for negative emotions (figure 3.8). Here again, the TomTom is associated with the highest scores (M = 72, SD = 19), with markedly lower mean ratings for the Blaupunkt (M = 52, SD = 21) and Mio (M = 46, SD = 24). Together, these differences are significant[14], $F(2, 36) = 5.35$, $p = .009$.

3.2.3. Discussion

Many of the user experience scales used in this experiment were found to be sensitive to differences between products within a single category (personal navigation devices) in a between-subject experiment designed to avoid explicit comparisons by the participants. In particular, the various navigation devices obtained significantly different scores on a modified version of the PrEmo questionnaire, showing it to be useful to measure emotional responses to interactive products.

Interestingly, some of these questionnaires (the "experience of meaning" questionnaire and PrEmo) were used previously in a distinct study with the same products but a completely different task, namely simply looking at the device and watching a video of someone else using it (Desmet & Schifferstein, 2010). The structure of the questionnaires was broadly similar in both cases but the pattern of self-reported emotions was completely different. This suggests that the differences observed here really do result from the interaction itself and not from some other properties of the products.

14 The data from one participant (using the Blaupunkt TravelPilot 500) were not included in the analysis because of a missing rating for "admiration".

3.3. Conclusion

In these two studies, two different emotion self-report questionnaires based on PrEmo were shown to be sensitive to differences between products both across categories (coffee machine and alarm clock) and within a category (personal navigation devices) across two different experimental designs. Interestingly, a comparison with an earlier study conducted with the same products suggests that these differences in self-reported experience are also specifically related to interaction with the product.

4. Moment-to-moment Measurement of Affect[1]

The various questionnaires used in chapter 3 have proven to be sensitive to the character of the interaction with consumer electronics or kitchen appliances but still only provide a single, punctual measure of the experience of each research participant. These data paint an overall picture of the emotions induced by an activity but they only represent the outcome of a particular sequence of use, i.e. the state of the person after interacting with a product, or perhaps an integrated evaluation based on several potentially contradictory responses elicited by specific features or attributes of the design.

Overall ratings of the experience therefore provide only limited insight into the course of the interaction and the designer's options to shape it. The premise of this thesis is that researchers and designers could benefit from information about the dynamics of the interaction – the ebb and flow of experience during the complex sequence of actions, sensations and decisions involved in the operation of sophisticated products – to determine which elements of the design contribute positively or negatively to the experience and how they combine to leave a lasting impression. Collecting moment-to-moment data on emotions as they unfold over time could help designers identify the key moments that define the user experience and the stages of the interaction they can act on to impact affective response.

Moving on to the study of these dynamics creates several important measurement challenges related to the specific nature of person-product interaction and the type of emotions that can be expected in that context. This chapter describes some of these challenges and presents an approach to tackle them. Finally, some key aspects or elements of this approach are examined in more detail.

4.1. Difficulties and trade-offs

Compared to the evaluation of responses to product appearance or sensory qualities, research on the experience of interaction with

1 This chapter is based on an article published in the proceedings of *Designing Pleasurable Products and Interfaces 2009,* subsequently selected for a forthcoming special issue.

products is fraught with difficulties. On the one hand, the intensity of the response to be expected is mild in most cases, complicating the use of some measurements (e.g. observation of facial expressions and psychophysiological recording). On the other hand, asking research participants to carry specialized equipment or to be actively involved in the measurement process (e.g. through self-report) can itself interfere with the experience. For example, obtaining repeated ratings even on very simple questionnaires can quickly become burdensome for test users and distract them from the other tasks at hand.

Some of these difficulties can be approached through a series of trade-offs that researchers have to make when devising a measurement procedure to study user experience or emotions in design.

4.1.1. Temporal resolution and richness

The first of these trade-offs lies between temporal resolution and richness in the content of the emotion measure. The more detail we seek on the temporal dynamics of emotion, the more difficult it is practically and theoretically to obtain data that goes beyond basic dimensions of affect, whether in the domain of self-reported feelings or behavioral and physiological processes. Conversely, at a more integrated level of analysis, measurement with detailed verbal scales and tools based on discrete emotions become more practicable and meaningful. There is a sort of continuum going from punctual or unique measurement to moment-to-moment recording over a period of time with a trade-off between the amount of information that can be extracted at each measurement point and the number of measurement points in the study.

At one end of this continuum, personality assessment or surveys often use very long questionnaires including several multi-item scales. In design-oriented emotion measurement, this type of techniques can be contemplated when respondents only have to report their feelings about a single product (e.g. Mooradian & Olver, 1997; Richins, 1997) or perhaps a handful of products, but long questionnaires become extremely demanding to the participants when they have to be administered repeatedly. Studies requiring repeated self-report over an extended period of time (e.g. diary studies about circadian mood cycles, Watson, Wiese, Vaidya & Tellegen, 1999) or for more than half a dozen stimuli (e.g. films as in Hewig et al., 2005 or pictures as in Mikels et al., 2005) therefore use either short questionnaires with only two or three dimensional scales or brief measures with single-item measures of categorical emotions.

At the other end of the continuum, research asking people to report more or less continuously their response to an ongoing stimulation (film, music, advertisement) are restricted to single measures assessing

one or two dimensions (Cowie et al., 2000; Geringer, Madsen & Gregory, 2004; Schubert, 1999). Even when the task only requires attending to some stimulus, it is simply impossible to consciously track more than a couple of attributes continuously. The only practical options to collect self-report data on more than two dimensions is to present the same stimulus several times to the same participants or to measure each dimension of interest with a different set of participants.

Several labels (e.g. "aesthetic response", "warmth") have been used in different fields to explain the measure to research participants, but evidence from music perception studies suggests that respondents might in fact understand many of them in a broadly similar way. Lychner (1998) found out that data collected by asking listeners to report the "felt emotional response" was very close to self-report about "aesthetic experience" or about an unspecified dimension anchored with the words "more" and "less" but not with "perceived tension".

This finding is broadly coherent with some of the models of emotion discussed in chapter 2 (see in particular section 2.1.3). The data collected by Lychner (1998) could thus be interpreted as reports of valence, pleasure or hedonic tone as there is considerable evidence that valence is the major dimension underlying many affective responses. Barrett (2006) reviews some of this evidence and articulates a view of valence as a "fundamental building block of emotional life", with discrete emotional states such as "anger" or "fear" as emergent properties in the perception of emotion. Similarly, Russell's (2003) influential notion of "core affect" is based on the idea that we constantly find ourselves in an affective state defined by two dimensions (valence and arousal) which provide the backdrop for more complex emotional phenomena, elaborated on the basis of this core affective state, its temporal dynamics and conscious and unconscious cognitive processes.

Under this model, the limited number of dimensions in moment-to-moment assessment is therefore not only a practical limitation due to the conscious involvement of the participant in the self-report process and attention or cognitive load coming with it but a fundamental property of affect. It is in fact not clear if we genuinely experience complex and elaborate discrete emotions every few minutes when using something but we certainly can tell at most times if we feel generally frustrated or satisfied. Rich measures of discrete emotions would thus be more meaningful for integrated judgments of a product or event as a whole whereas dimensional, and especially valence-based, formats would be more appropriate for continuous or frequent moment-to-moment measurement of experience. Indeed, research with continuous measures that do not involve self-report also has difficulties differentiating affective states beyond basic dimensions like valence and arousal (Larsen, Berntson, Poehlmann, Ito & Cacioppo, 2008).

4.1.2. Level of interference and distance from interaction

The second trade-off faced by design-oriented researchers is between the level of interference in the situation and the distance between the original activity and the measurement itself.

At a very general level, this trade-off surfaces in the choice between market research surveys and organized product tests. In surveys about consumption experiences (e.g. Richins, 1997) or long-term retrospective studies (Karapanos, Zimmerman, Forlizzi & Martens, 2010), there is virtually no interference with the interaction itself: Participants are invited to respond based on past usage of a product they chose themselves before the start of the study. Before recruiting the participants and asking them specific questions, the researchers do not have any influence on the respondents' activity or the products they use in their daily lives. The distance, however, is high: Ratings rely on the memory of events sometimes far removed temporally or geographically from the moment the data are collected.

Lab or field-based product tests represent another trade-off between interference and distance: Researchers interfere heavily with the participants' usage pattern by prompting them to interact with a specific product and defining the tasks to carry out but it becomes possible to collect data about the user experience associated with a well-defined interaction sequence, immediately during the test or shortly afterwards.

Even in experimental research, the choice between measurement procedures involves a trade-off between the level of interference and the distance between the interaction and the data collection. Thus retrospective self-report lets participants interact relatively freely with a product within the confines of the lab whereas repetitively prompting them to provide concurrent self-report during use interrupts the activity and threatens to disrupt the flow of experience. Moment-to-moment affect ratings as practiced in fields like music or advertisement research represent an extreme form of trade-off: Data are collected instantaneously as the experiment unfolds but the measurement places a very heavy burden on the participants, requiring to constantly monitor and report their own feelings. The techniques used in these fields can only be applied when the experimental stimuli can be processed "passively" without manipulating or interacting with any other device than the data collection device itself. Even then, it is difficult to believe that concurrent self-report does not affect sensory or affective processes and there is a risk that research participants incur extra attentional or cognitive load that could fundamentally interfere with the perceptual processes themselves. When dealing with interactive artifacts rather than media stimuli, participants need both

to be able to attend to other goals than simply rating something and to have their hands free to operate the product.

Techniques such as psychophysiological measurement or automatic facial expression monitoring offer the promise of practically continuous online assessment of emotional responses without requiring any active involvement of research participants. In this case, the interference with the activity comes from cumbersome equipment and restrictions to participants' movement. In some extreme cases (e.g. brain imaging with functional magnetic resonance imaging) subjects have to lie still in a cramped space inside a noisy machine but for some other measures, progress in ambulatory physiological measurement and wearable sensors greatly reduced these constraints. For example, after a short adjustment period, modern electrocardiography equipment is barely noticeable and can be worn for hours. Affective computing seeks to leverage these progresses to achieve continuous detection of emotions without any active involvement of the person experiencing them and could provide a way out of the interference/distance conundrum.

4.2. General approach

Two fundamental ideas guided the design of the measurement technique presented in this chapter: The multi-componential nature of emotion (see chapter 2) and the need to avoid disrupting the flow of experience during interaction. Adopting a multi-componential view of emotion naturally led to the exploration of measurement based on various components, such as physiological recording and expressive behavior. But it also means that conscious feelings are understood as a key part of emotions elicited by products. Self-report is therefore relevant on theoretical grounds and not merely an inferior approach that subsists because of the technical difficulties associated with other forms of measurement.

Figure 4.1. Approach to the measurement of the dynamics of emotion in person-product interaction[2] A: Physiological recording equipment can be attached for ambulatory measurement during the test. B: The test participant interacts with the product freely while being filmed. C: Video of the product test is presented immediately to collect emotion ratings

The approach developed in this thesis is built around video-supported retrospective measurement ("self-confrontation") to collect moment-to-moment ratings of emotional experience without requiring active involvement of the research participants in the measurement process at the time they are using the product (see typical procedure in figure 4.1). Additionally, other measures can be collected during the test (traditional questionnaires, physiological recording) and, depending on the setup, a video feed can be used to code facial behavior.

The core principle of self-confrontation is to first let participants complete their task without being interrupted. They are videotaped while using the product and report their feelings immediately afterwards using the video to support their memory of the activity and of their experience of the interaction with the product. This technique can be seen as a way to strike a balance between staying close to the activity and avoiding to interfere with it. Self-confrontation combines a form of retrospective self-report, limiting interference with the person-product interaction, with the use of video as a recall cue to collect detailed information about its dynamics. Depending on the research questions or the stage of the design process, it can be adapted by using different data collection approaches: open-ended questioning or more structured questionnaires. Self-confrontation studies can therefore vary in response format.

4.3. Aspects of the procedure

The approach sketched above includes several phases or stages, starting with the product usage phase itself followed by the self-confrontation phase during which participants provide moment-to-moment ratings of their experience. Such a complex technique raises

2 I am thankful to Anna Fenko for serving as a model and to Pieter Desmet for preparing this illustration.

a number of questions on the details of the procedure. The rest of the chapter is devoted to a detailed discussion of some of these aspects, providing a rationale for some of the important decisions made when designing this approach.

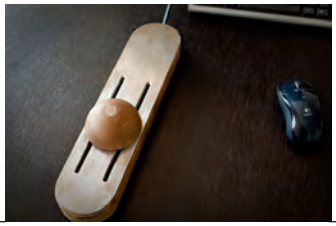
4.3.1. Self-confrontation

The self-confrontation technique is the main element of the measurement procedure and it is instrumental in collecting self-report data without interfering with the flow of experience as users interact with the product. The basic principle is that participants are filmed while interacting with each other or with artifacts. They are then asked to report their feelings while watching a video of the interaction, immediately after it ended. The same technique can also be used to collect qualitative data about the interaction, probing for more information on key events revealed by the ratings. The following pictures[3] illustrate the main steps of the procedure.

3 I am thankful to Pieter Desmet for serving as a model and to Chajoong Kim for taking and processing the pictures.

Research participants are first filmed as they interact with a product. The angle varies depending on the practical constraints of each study but is chosen to capture a subjective view of the situation avoiding any third-person shot of faces.

Immediately after the interaction, simple valence ratings are collected by showing the video to the participants and asking them to report how good or bad their feelings were.



A purpose-built device, the emotion slider (see chapter 6), is used to record the moment-to-moment ratings.

Visual feedback can be displayed beside the video as the ratings are collected.

The valence ratings can be immediately displayed and used during an interview to collect qualitative data on the participants' interpretation of their feelings.

Ratings (displayed under the video) are clickable and can be used to navigate through the video.

Self-confrontation is not altogether a new technique. In fact it has already been used in human-computer interaction research before, for example to collect open-ended qualitative data on the thought process of website users (Lim, 2002) or ratings of cognitive workload from naval operators (Neerincx, Kennedie, Grootjen & Grootjen 2009), but also in research about the affective aspects of user experience (Cahour et al., 2005; Krone, Hamborg & Gediga, 2002). The originality of the present work is that it extends the technique to the collection of quantitative data and to the moment-to-moment measurement of emotional valence.

In self-confrontation, the purpose of the video is to support the self-report, helping the participants to remember their experience and allowing them to report more accurately on the time course of the interaction. Self-confrontation can therefore be contrasted with concurrent self-report on the one hand and with purely retrospective self-report on the other hand.

Delayed or retrospective self-report can in principle allow the collection of meaningful data on the emotional experience while limiting interference with the interaction as it happens. Relying solely on the participants' memory and ability to recall a complex sequence of events freely however provides only limited insight into the course of the interaction and risks introducing additional biases in the self-report. For example, participants are likely to remember only a few salient details or have a distorted view of the chronological sequence of events. The video should serve as a cue to limit these biases and support self-report during the self-confrontation phase. Even if they are based on memory, the ratings are closely linked to the events in the interaction and follow the actual time course of the sequence.

Self-confrontation could therefore improve the validity of the data compared to a classic retrospective assessment and provide valuable data to design researchers and practitioners. However, it is quite new and has not been used very often in quantitative or affect-oriented research, leaving many questions about the technique and the details of the procedure open.

On a practical level, an important question pertains to the cues that best help the participants to recall their experience. Different cues could be used with the self-confrontation procedure, from screen captures (for software products) to various types of videos differing by the camera angle, presence or absence of sound, etc. Anecdotal evidence suggests that seeing one's own face or hearing one's own voice is a rather unusual experience that can generate surprise and embarrassment, potentially prompting participants to focus more on their situation during the self-confrontation phase than on their experience at the time of the interaction. Conceivably, this could foster a more reflective perspective and make the presence of an observer even more salient.

These considerations justified the choice of a quasi first-person view, with the camera positioned right behind the research participants, filming them from the side. Their hands and body are therefore sometimes visible on the video but the angle corresponds broadly to the view one would have had when using the product being tested. Ensuring that any computer or other screen is visible and legible on the video should also be a concern when planning a self-confrontation study.

Obviously, such a set-up does constrain the type of tasks and interaction that can be studied but it is by no means strictly restricted to seated, lab-based tests, as illustrated by the two studies described in chapter 5. Alternatively, a small camera mounted on a light helmet or pair of glasses could provide an even more compelling subjective view while completely freeing the participants' movements.

There is no strong empirical or theoretical basis to decide on the presence of sound but it is often necessary to include it on practical grounds, as it is an important feedback channel in the design of many products, including several of those used in the present research (alarm clock, personal navigation devices).

It also seems important to ensure that self-confrontation ratings are collected quickly after each interaction sequence, while the memories are still fresh[4]. Small digital cameras give researchers some flexibility in the setup and allow a quick transfer of the resulting video to a computer. Custom-software was developed to collect the actual ratings and be able to synchronize the data with the timeline of the video.

4.3.2. Moment-to-moment self-report with the emotion slider

Another set of questions pertains to the format and content of the self-report data themselves. A straightforward solution would be to repeatedly prompt research participants to report their feelings with a (brief) questionnaire (Lee & Jeong, 2006), perhaps one of the emotion self-report scales described in chapter 2. Design-oriented researchers tend to use idiosyncratic scales addressing perceived deficiencies

4 But see Redelmeier & Kahneman (1996) for a different view, in the context of pain research. Comparing different forms of self-report during a painful surgical procedure, Redelmeier and Kahneman found that patients formed a judgment about the overall level of pain immediately at the end of the procedure and that this judgment did not reflect the average level of pain reported during the procedure. Interestingly, this judgment also remained stable over a month. In short, retrospective self-report provided a distorted view of the pain experienced during the procedure, independently of the time elapsed since.

of general emotion questionnaires, with single-item descriptors of emotions chosen on the basis of researchers' best guest and of the focus of the study at hand. Such *ad hoc* measures can however be detrimental to the comparability of the results and the development of the field and could advantageously be replaced with standardized measurements developed for product evaluation.

In any case, as noted above in section 4.1.1, repetitive self-report with lengthy scales can become burdensome for the participants. An alternative approach is to use a simple dimensional moment-to-moment self-report similar to the measures used in music or advertisement research examined in chapter 2. Even then, the specific content of the self-report has to be considered carefully. In keeping with the theoretical literature on the importance of valence as a fundamental dimension of affect, the instructions used for the self-confrontation studies in chapter 5 describe the response in very general terms and ask participants to provide moment-to-moment ratings of how good or bad they felt during the interaction. The software developed for these studies also enables the researcher to present these ratings immediately back to the user. The valence ratings can then be used as a starting point in the discussion with test participants in an open-ended interview to collect more interpretive data about their feelings.

Moment-to-moment self-report also typically relies on custom input devices such as dials or button boxes. Since research participants have to provide online ratings while attending to something else, the interface used to collect these ratings is both more complex and more sensitive than it would be for a regular questionnaire. The shape and physical characteristics of the self-report device could therefore also have some influence on the data obtained but little research seems to be available beyond the discussion of the instruction and labels used to describe the response of interest. A basic methodological precaution, common in some fields, such as music perception research, is to invert the self-report scales for half of the participants, for example by switching the positions used to report positive feelings and negative feelings. This strategy can in principle mitigate a systematic bias in favor of a particular movement or direction but it does not prevent a confusing device to cause random errors or hesitations.

There is in fact a growing literature on the congruence between instrumental behavior and affect, and basic approach/avoidance tendencies are often mentioned as one of the key components of emotion. Nonetheless, it seems that little attention has been paid to the type of motor responses required from participants in user experience or media psychology research. The emotion slider, described in more details in chapter 6, was developed based on this literature and on the principles of tangible interaction to facilitate affective self-report during self-confrontation.

The shape and mechanical properties of the emotion slider have

been designed to maximize the congruence between the physical response and the content of the feelings being reported. The research reported in chapter 6 does support the hypothesis that the tangible characteristics of the slider provide an intuitive mapping with valence or emotion intensity and could therefore make visual feedback redundant.

4.3.3. Multi-componential measurement

The last aspect of the procedure that deserves further discussion is the role of other components of emotion than feelings and subjective experience in the approach presented here. The procedure does allow for the collection of other physiological and behavioral data and the lack of interruption during the activity itself would certainly benefit these kinds of measurement. Chapter 9 discusses a number of difficulties with this type of data but, as noted above, using ambulatory measurement equipment or wearable sensors for electrocardiography during a product test is reasonably easy on a practical level.

Some other signals do create some specific logistical challenges in interactive settings. Two of them, skin conductance and facial behavior, will be discussed in a little more detail. For anatomical reasons, reference texts on skin conductance strongly recommend placing sensors on the palm of the hand, which is obviously not possible when research participants have to move their hands and manipulate objects. Some researchers dealt with this difficulty by attaching the electrodes to an arm or a foot but the consequences for the quality of the measurement are unclear.

Facial expression can also be recorded easily, either with surface electrodes (electromyography) or through direct observation. Each approach has its own advantages and disadvantages. Electromyography is more sensitive but facial electrodes are slightly obtrusive and more annoying than electrocardiography sensors. Observation of visible facial behavior requires an extra camera with a clean frontal shot of the head, further restricting the participants' movements.

Both facial expression and autonomic physiology have a clear advantage for the moment-to-moment assessment of the dynamics of emotion; these data are naturally continuous and can be sampled with a high frequency, potentially offering a very high temporal resolution, at least at the level of the physiological signal. Analysis and interpretation however only rarely realize this potential. As noted in chapter 2, nearly all published studies average all physiological data collected during each experimental condition, aggregating changes from baseline across several trials. Other analysis strategies need to be developed and applied to user experience research for these techniques to be useful to the study of emotion dynamics in design.

The multi-componential view of emotion was also one of the starting points of this work and provided a structure for the review in chapter 2 or the discussion of validity in chapter 9. Still, all our experimentations with physiological measurement (both autonomic physiology and facial electromyography) have been unsuccessful and these techniques were not included in the empirical studies reported later. Chapter 9 does however discuss a number of theoretical and methodological issues related to the use of this type of measurement in design-related research.

Lastly, there is a big discrepancy in the way physiological measurement is understood and embedded in research in different fields. In psychophysiology or neuroscience, bodily changes and interactions between these changes and psychological processes are of great empirical and theoretical interest in and of themselves but self-report is routinely integrated in experimental protocols and often serves, directly or indirectly, as a point of reference to index relevant psychological processes. In some applied fields however, there is a strong emphasis on avoiding any form of self-report either for practical (e.g. achieving completely implicit interaction in affective computing) or methodological reasons (e.g. the belief that psychophysiological measures are better or less susceptible to some biases). This emphasis often leads to a lot of theoretical confusion and disappointing results[5].

Instead of looking at physiological data as objective measures of emotion bound to replace self-report in the near future, it could be useful to consider ways to combine them with other approaches. These data could for example be used to identify key episodes during the use of a system. It would then be possible to ask users if they indeed experienced stronger feelings at that time and to probe further about the content of these feelings, either online with some form of short questionnaire or offline during self-confrontation. Spurious detection of emotion (false positives), lack of specificity or ambiguity could be compensated by the self-report data, while the other streams of data could help the researcher to decide at what time to probe for more detailed self-report and increase the validity of the results.

4.4. Conclusion

This chapter detailed the specific difficulties that researchers face when they want to assess the dynamics of affect in a design context.

5 In fact, avoiding self-report is rarely possible in practice, but this inclination is evident in sweeping proclamations about the value of psychophysiological measurement in introductions and conclusions.

Two major trade-offs – between temporal resolution and richness and between interference and distance from the interaction – were identified and an approach to the moment-to-moment assessment of emotion during person-product interaction was sketched. This approach represents an attempt at striking a balance between the different constraints.

Thus, unidimensional self-report was adopted as a way to maximize the temporal resolution and allow practically continuous measurement of affect. However, to keep the participant free to interact naturally with the products being tested, these moment-to-moment self-report data are not collected concurrently but right after completing the test, using self-confrontation to stay as close as possible to the temporal dynamics of the person-product interaction.

These choices are based on our best judgment but also in no small parts on practical contingencies. Other choices could be made based in particular on the specifics of the products studied and on the objectives of the researchers. It is to be hoped that the research reported here and future studies using self-confrontation can inform these choices.

5. Self-confrontation

The core of the moment-to-moment emotion measurement procedure described in chapter 4 is the self-confrontation technique. By combining video recording and moment-to-moment rating, it aims at collecting self-report data about a research participant's feelings, time-locked to the interaction but without interfering with it. Using such a new and complex approach obviously raises a number of important questions about the data collected and their interpretation, some of them discussed in chapters 4 and 8.

The most basic of these questions is whether or not the data really reflect product-related differences in experience. A straightforward way to establish that self-confrontation ratings can be used to compare different designs with similar function is to ask users to interact with products expected to elicit different experiences and compare the resulting data. If there are independent empirical or theoretical reasons to believe that a given product should elicit more positive feelings than another one, measures collected during interaction with the former should yield a more positive score than measures collected during interaction with the latter.

Self-confrontation was therefore used in two studies with products that were expected to generate very different experiences. To some extent, the contrast between the stimuli selected makes these tests something of a "toy" situation. Indeed, the focus of these experiments was not primarily on learning something new about the products but rather to establish a link between the differences in the products and the scores collected during self-confrontation.

To assess the viability of the approach, several experiments were conducted with the procedure. The first experiment used an early prototype of the self-confrontation software and vases and cameras as stimuli. The second experiment was conducted using the emotion slider, a purpose-built input device described in more detail in chapter 6, and a new version of the software. Additionally, the moment-to-moment emotion self-report were compared with post-use ratings of user experience to assess the relationship between self-confrontation and other methods and to illustrate the potential of the technique to investigate how ongoing experiences are integrated to form an overall judgment of a product.

5.1. Experiment 1: Vase and camera[1]

The fist experiment was the first step towards using self-confrontation in a quantitative fashion to measure emotion in a product-usage context (see also chapter 4). Participants were asked to complete a task involving several products: arranging flowers in a vase and taking a picture of it. The experiment followed a within-subject design and all participants were exposed to all products included in the study. Right after completing the task, the participants reported their feelings while watching a video of the interaction. Data collection proceeded using an early version of the self-confrontation software, operated with the keyboard. Participants could therefore rate discrete events with a dichotomous response format (positive or negative feeling). Additionally, post-test interviews provided an assessment of the face validity of self-confrontation as an emotion measurement.

5.1.1. Material and methods

The main stimuli were two different vases, selected on the basis of the emotional responses that they were expected to elicit during use (figure 5.1). One of the vases was a small cubic vase made of thick glass. The 55 centimetre-long flowers did not fit nicely in it and even tended to fall down, hence making the experience with this vase a rather frustrating one. The other one was a tall, translucent plastic vase looking like a glass vase. It was therefore much lighter to lift as could be expected from its appearance and was predicted to be surprising and fun to use, as shown by previous research with the same product (Ludden, Schifferstein, & Hekkert, 2006).

Figure 5.1. *Stimuli used in experiment 1: frustrating (left) and surprising (right) vases.*

1 This section is based on an article published in the proceedings of *Design and emotion 2006*.

Participants (N=25, 14 women, 11 men) were students at the Industrial Design faculty of Delft University of Technology. They were approached during the breaks in the free-time area of the building and asked if they would like to participate in a test involving a "new approach to get feedback about peoples' feelings when using products". They were paid a small compensation fee to participate.

Participants were asked to follow a scenario to "test their new digital camera". They had to "make a nice composition" with some artificial flowers and a vase. Then, they took a picture of it and downloaded this picture on a computer. While such a complex scenario complicates the interpretation of the results, creating a situation that would come sufficiently close to actual product usage to elicit comparable emotions is necessary to assess the relevance of the technique for product evaluations and research on person-product interaction. Yielding useful data in this type of relatively uncontrolled situations is in fact a *sine qua non* for a design-oriented tool. Additionally, the scenario added a goal-directed aspect to the task by inviting participants to make a nice composition to be able to test the digital camera. This task is in line with appraisal theories of emotion (Scherer, Schorr, & Johnstone, 2001), which predict that emotions arise – among other situations – when an individual is faced with goal-conducive (or, on the contrary, hindering) events.

To support the story and prevent the participants from focusing solely on the vase, the experiment also involved two different digital cameras. While the order of presentation of the vase and camera could not be counterbalanced without making the scenario meaningless, the product combinations were randomized (i.e. some participants used camera A with vase A first, some started with camera A and vase B, some had camera B with vase B first and so on, see table 5.1).

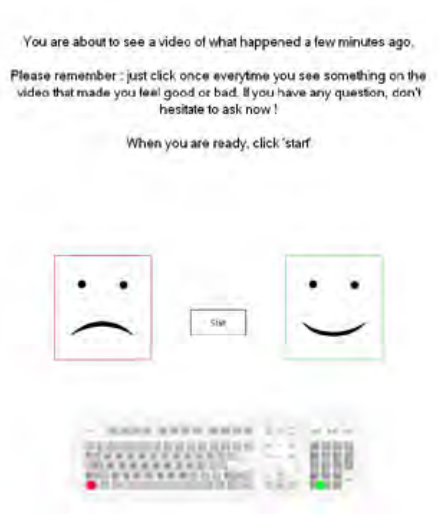
Table 5.1. *Overview of experimental design*

Vase used first	Camera used first	N
Frustrating vase	Canon	7
	Fuji	6
Surprising vase	Canon	6
	Fuji	6

The test took place individually in a usability lab-type facility. After a short introduction, the participants had to read and approve a consent form. They were then seated at a computer and presented with an on-screen demo of the rating procedure they were to use after completing the tasks together with some explanation about the course of the test.

A scenario card was handed out to them and they were asked to read it and wait for the moderator to be ready to record the test

before starting. The field of the video camera included the table, vase, flowers, camera and computer the participants had to use. The setup resulted in a ¾ shot of the participants, from the side. When they finished carrying out the task, the participants had to wait between 1 and 3 min for the video to be converted and saved on the computer



before they could start the self-confrontation. For technical reasons, this delay depended on the time spent carrying out the task.

Figure 5.2. *Instruction screen for the self-confrontation procedure.*

The self-confrontation itself took place in the same room, on the computer used during the introduction. The software was developed specifically for this test and started with a screen reminding the participants of the instructions given to them at the beginning and inviting them to ask any question they might have before starting the self-confrontation (figure 5.2). After pressing the "start" button the video appeared and participants could report experiencing a positive or negative feeling at any time until the end of the video. To do so, they had to press one of two buttons (the left "Ctrl" key or the "Enter" key from the numeric keypad). These buttons were situated at opposite ends of the keyboard and were to be operated with a different hand each.

Little coloured stickers on the keyboard itself linked the buttons to the two faces on the screen, which were themselves contained in assorted coloured frames. As in other tools like the SAM (Bradley, & Lang, 1994) or the 2-Dimensional Emotion Space (Schubert, 1999), a smiling face stood for positive valence ("feeling good about something") while lip corners pulled downwards represented negative valence (see figure 5.2).

After reaching the end of the video, the software automatically stopped and invited the participant to turn to the moderator. A short interview followed, with four open questions: about feelings during the test, the products in general and then about the feelings and opinions associated with the camera and the vase in particular, in that order. At the end of this interview the participants were handed out the second scenario card and went through the same procedure with the vase and the camera they did not use yet.

After both tasks and self-confrontation sessions were completed, a debriefing interview concluded the test. The moderator queried about the participants' opinion about the software, if they felt confident they could remember their feelings and finally if they thought this procedure would provide a good way to get feedback on people's feelings with products.

5.1.2. Results

The data collected were in the form of a list of reports with, for each press of a button, the amount of time since the beginning of the video and the valence (positive or negative) of the experienced feeling. The data from two participants could not be included in the analysis because both of them chose not to put the flowers in the vase in one of the two trials. The number of reports per trial varied widely (M = 12.72, SD = 8.32) for a total of 636 data points. Timing of key events in the interaction (first contact with the vase, first attempt to put the flowers in the vase, first contact with the camera) was coded from the video.

Based on these events, all reports recorded in the 8s following the first attempt to put the flowers in the vase were extracted. The 8s delay was chosen somewhat arbitrarily to represent the users first experience in using the vase, the primary outcome in this test. Using a fixed "window" (as opposed to the full episode) seemed a simple and efficient way to avoid biasing the results by the time each participant took to complete this subtask. In any case, participants rarely used less than 8s to complete this part of the scenario.

All reports were then added, giving the weight -1 to negative reports and +1 to positive feelings, yielding two summary ratings (one for each trial, i.e. each vase) per participant. It must be noted that this computation precludes any distinction between, for example, "no feelings" (i.e. no report at all) and multiple reports adding up to 0 (i.e. exactly the same number of positive and negative reports).

Still, this simple computation gives an overview of the type of feelings that dominated in the experience of the user. Altogether, participants reported between 0 and 3 feeling episodes per trial for a total of 22 events for vase 1 and 19 for vase 2.

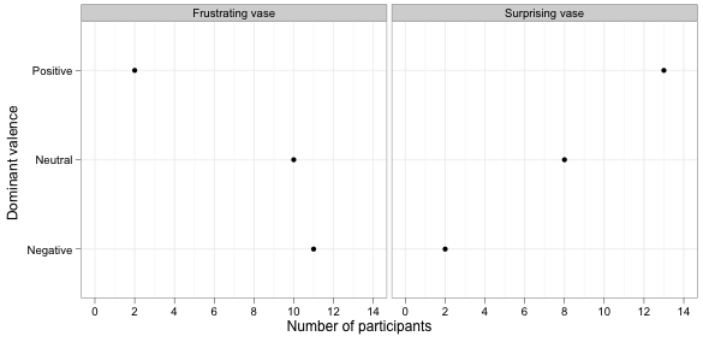


Figure 5.3. *Dominant valence for each participant with the frustrating (left) and surprising (right) vases. Experience is coded as positive for participants reporting more positive than negative events and vice versa. A "neutral" experience corresponds to an equal number of positive and negative events or to no events reported at all.*

As shown in figure 5.3, 48% of the participants (i.e. 11 out of 23) reported more negative feelings than positive feelings after using the vase predicted to be frustrating whereas the proportion was inverted for the surprising vase, with 56% reporting more positive than negative feelings (neutral responses were respectively 44% and 35%). A sign test confirmed that the difference was significant (N = 19, 4 ties, p = .001). Even if the contrast is obvious from figure 5.3, the sign test provides a simple way to test the significance of this difference, with minimal assumptions. Since no specific predictions were made regarding the experience of the camera, no such analysis was performed for the interaction with the camera.

The same scoring procedure was also applied to the whole dataset (i.e. not limited to the 8s window following contact with the vase). Difference was much less marked in this case with 35% negative and 61% positive reports for the surprising vase and 39% negative and 48% positive feelings for the frustrating vase. Unlike the test comparing events recorded right after interacting the vase, a sign test conducted on the sum of all events was not significant.

5.1.3. Discussion

The results generally support the main hypothesis that meaningful data about user experience can be collected through a self-confrontation procedure. Moment-to-moment measurement, together with the video, allowed fine-grained analysis of key episodes in this relatively simple scenario. Interestingly, the different feelings elicited by the contact with the vase are not apparent in an analysis including the whole sequence of interaction and would probably be hidden in

analyses of aggregated data or overall retrospective assessment with a classical self-report instrument.

Informal analysis of the interviews carried out after the test also suggested that the method enjoyed a relatively high acceptance from the participants. Nearly all of them were confident they could adequately remember and report about their experience. However, they were also keen to stress several conditions under which they felt this technique should be used. Among them is the very short delay between the actual test/interaction with the product and the self-confrontation.

Many participants also highlighted the importance of the post-test interview to articulate their feelings in more details and provide explanations regarding the reasons they were feeling in a particular way. As expected, many participants also resented the limitation of the self-report to just two possible emotional states ("positive feeling" and "negative feeling") and expressed the need to be able to report intermediate states and/or qualitatively different feelings.

5.2. Experiment 2: Personal navigation devices

The vase-and-camera study represented the first use of self-confrontation to assess emotions during interaction with products but it suffered from a number of weaknesses. To test the procedure in another context and address some of these limitations, self-confrontation was also included in the personal navigation device test already described in chapter 3 (see section 3.2). The main differences between the vase-and-camera study and the navigation device study are the experimental design and the type of moment-to-moment self-report data collected during self-confrontation.

The experimental design selected for the vase-and-camera study meant that each participant saw all the products tested. Such a within-subject design is very popular as it reduces the number of participants needed and mechanically controls many potential confounding variables and individual differences, therefore being more sensitive. It does however suffer from a number of disadvantages including fatigue and learning effects but also the potential to increase demand characteristic effects (Orne, 1962; see also chapter 8) by making the researcher's interest and hypothesis manifest to the participants. Indeed, showing two products one after the other strongly suggests that a difference is expected and could led the participants to consciously or unconsciously alter their behaviour in response to this expectation. Having each participant use only one product of course does not completely remove demand characteristics from

the experimental situation but it does deemphasize the differences expected by the experimenter and generally provide a more stringent test of the discriminatory power of the measurement used. It is therefore important to test the self-confrontation procedure with different designs.

The second major difference between experiment 1 and experiment 2 was the format of the self-report. This time, the moment-to-moment ratings during self-confrontation were practically continuous, using the emotion slider, a device designed to allow participants to report their feelings at any time[2]. The procedure was also extended to provide participants with a way to elaborate on their ratings in a post-self-confrontation interview.

Finally, experiment 2 also included several post-exposure questionnaires about emotion, perceived usability and user experience, detailed in chapter 3, section 3,2.1. These measures are used here to provide a comparison point and evaluate the value of the information collected during self-confrontation.

5.2.1. Material and methods

The procedure and material used in this experiment are described in detail in chapter 3. In short, 40 participants were given one of three personal navigation devices (see figure 3.4) and asked to drive to pre-defined locations in Delft. Specifically, the participants first had to follow driving instructions to reach two pre-programmed points (task 1), to enter the address of the university using an instruction sheet and to drive back to the university following the device's instructions (task 2). At the end of the drive, they came to a lab and completed various user experience questionnaires before going through the self-confrontation procedure.

A printed leaflet explaining the procedure was given to them while one of the experimenters transferred the video from the drive (example in figure 5.4). Participants were instructed to report positive feeling by "pushing the handle toward the screen" and negative feelings by "pulling the handle away from the screen". After reading these instructions, they watched the video of the drive while reporting their feelings with the emotion slider. This self-confrontation was followed by an interview.

5.2.2. Results

The post-use ratings on various user experience scales are detailed

2 See chapter 6 for more details on this device and its development.

in chapter 3. Importantly for the comparison with the moment-to-moment emotion ratings, there was a significant difference in the overall pleasantness or valence of the experience as indexed by PrEmo ratings collected after completing the two driving tasks.

Figure 5.4. *Snapshot from one of the videos (mirrors obscured for privacy reasons).*

As shown in figure 5.5, the mean score for the TomTom personal navigation device is the highest (M = 72, SD = 19), with markedly lower ratings for the Blaupunkt (M = 52, SD = 21) and Mio (M = 46, SD = 24). An omnibus test confirms that the various devices elicited significantly different retrospective emotion self-reports, $F(2, 36) = 5.35$, $p = .009$.

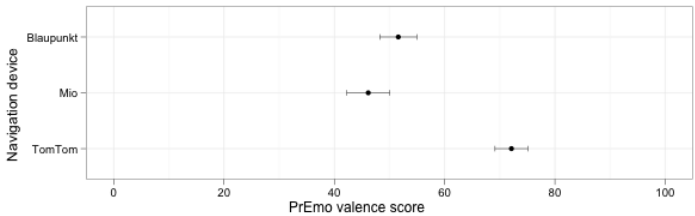


Figure 5.5. *Mean retrospective emotion ratings (error bars: standard error of the mean; graph reproduced from figure 3.8).*

A look at the raw self-confrontation ratings at the participants' level reveals huge individual differences, not only in the overall valence of the experience but also apparently in response style and in the way to report feelings. Figures 5.6.1 to 5.6.7 provide examples of individual ratings.

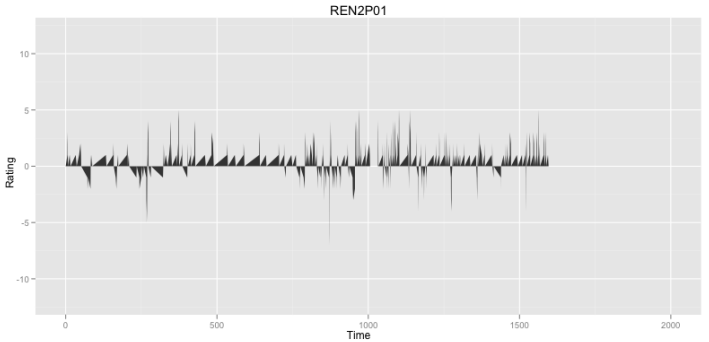


Figure 5.6.1. *Self-confrontation rating from participant 1 (time in s). This participant only reports brief punctual experiences and uses less than about a third of the amplitude available to report feelings.*

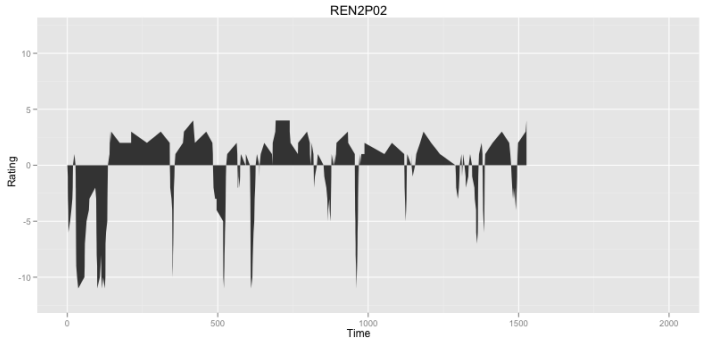


Figure 5.6.2. *Self-confrontation rating from participant 2 (time in s). This participant does not report as many changes in feelings as the previous one. Self-reported positive experiences are sustained for several minutes whereas negative experiences are short burst of negative feelings Ratings are also asymmetric reaching much further on the negative than on the positive side.*

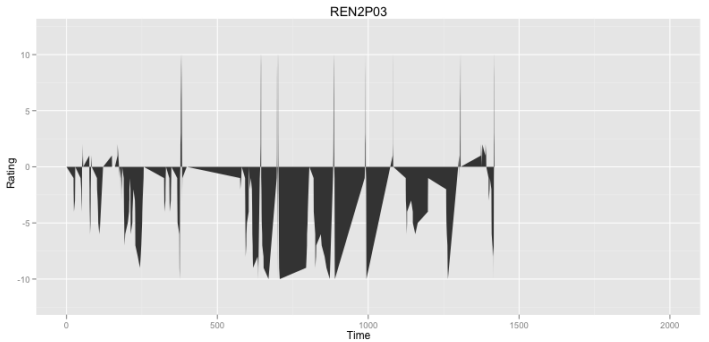


Figure 5.6.3. *Self-confrontation rating from participant 3 (time in s). Self-reported experience is almost exclusively negative with brief episodes of positive feelings. Ratings use the whole amplitude available with little nuance between the extreme positions.*

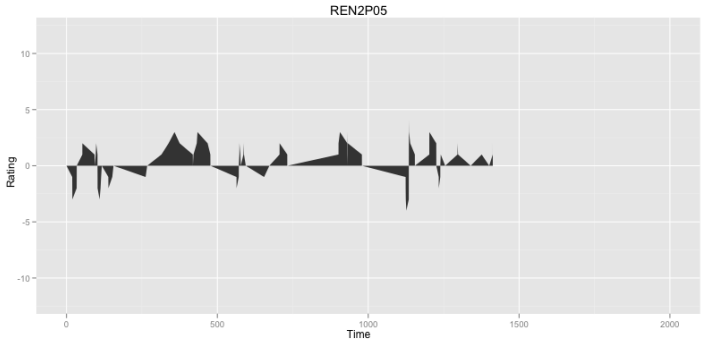


Figure 5.6.4. *Self-confrontation rating from participant 5 (time in s). Self-reported experience is changing slowly, with alternating phases of positive and negative affect using only a small fraction of the available amplitude.*

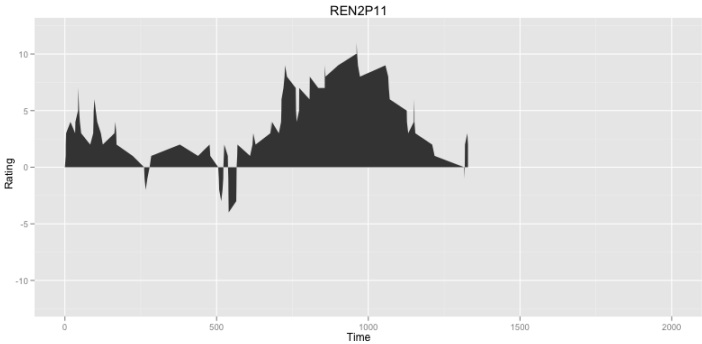


Figure 5.6.5. *Self-confrontation rating from participant 11 (time in s). This participant reported almost exclusively positive experience, using the full amplitude and nuances available on this half of the self-report device.*

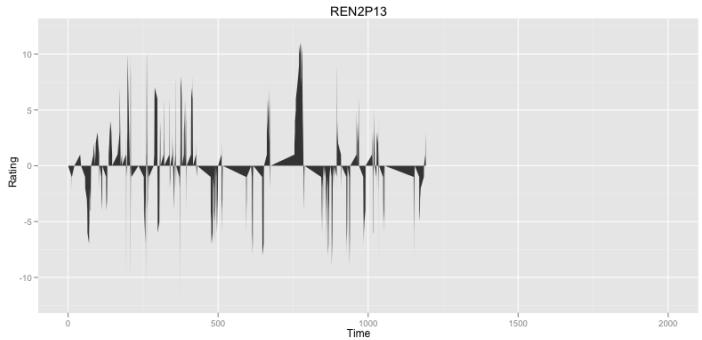


Figure 5.6.6. *Self-confrontation rating from participant 13 (time in s). This participant reports brief spikes of experience, using most of the available amplitude, in both directions.*

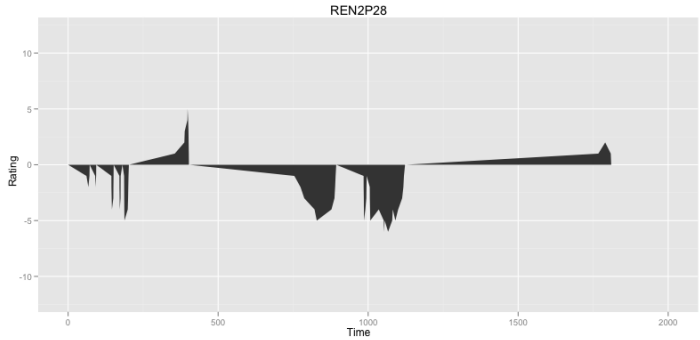


Figure 5.6.7. *Self-confrontation rating from participant 28 (time in s). This participant reports only a handful of key episodes lasting several minutes. The drive also took longer than for all other participants whose ratings are represented above.*

Several issues complicate the analysis of these moment-to-moment valence self-reports. The total time-on-task (i.e. the time spent driving) varied from participant to participant, from 20 to 35 min. Simply summing time-locked ratings across participants, as is often done with moment-to-moment data related to stimuli with a precise duration (films, musical excerpts, commercials) was therefore not an option. This problem stems directly from the interactive nature of the activity and the constraints of a field study. The time needed to complete such a task can't be fixed in advance and depends on several factors including the users (driving style and abilities, errors), products (guiding effectiveness of the personal navigation device) and extrinsic variables (in this case traffic and weather). Additionally, differences in total time reflect a myriad of smaller differences (staying at a particular red light, missing a turn, etc.) and the time scale for a given participant cannot be assumed to be linearly related to the time scale for any other participant.

A "quick-and-dirty" approach was adopted to deal with this problem. First the original data was resampled at 1Hz and smoothed with a 60s moving average. The timing of the beginning and completion of each task was then manually coded from the videos and used to "stretch" or "compress" the time to roughly align all series of self-confrontation ratings. Of course, different events might have happened to different participants at the same time

The variability highlighted above makes any kind of aggregation somewhat questionable. Still, a visual comparison between aggregated raw scores (see figure 5.7 for an example) and sums of scores normalized within participants did not seem to produce any major alteration. The rest of the analysis is therefore based on unstandardized scores, averaged across participants. These average ratings therefore represent the valence of the emotion at any given time, much in the

same way that mean scores on a post-test self-report scale represents average experience over the whole experiment. As described below, these aggregated ratings did reveal meaningful patterns of experience, vindicating this analysis strategy.

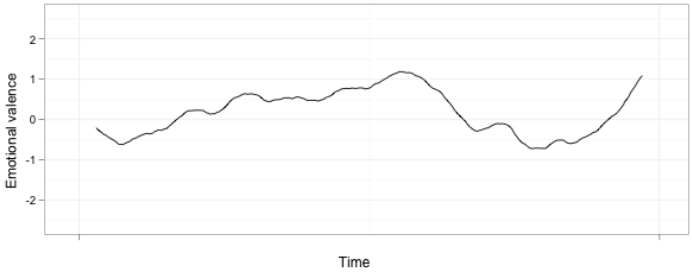


Figure 5.7. *Aggregated valence ratings for the first task (driving to two pre-programmed points).*

Figure 5.7 presents the average valence ratings across all devices during the first task (driving to two pre-programmed locations). An interesting pattern emerges across all three devices, revealing several easily interpretable phases. First, participants apparently went through a discovery and learning phase associated with neutral to mildly negative experience. Next, participants report mostly positive feelings, which correspond to a relatively easy part of the route that was followed without problems by most participants. After that, the ratings drop before rising again as participants attain their goal. The negative valence of the feelings associated with the last section of the route can be readily interpreted as a result of the difficult topography of the residential part of town where the objective was located and to the poor usability of most products used in the study, letting participants unable to understand the driving instructions provided by the navigation devices.

Interestingly, there is a clear interaction between the device used and the emotions experienced in each phase (figure 5.8). Whereas all three devices start more or less on an equal footing, self-reported experience improves rapidly for one of the navigation devices (TomTom XL), more slowly for another (Mio Moov) and barely, if at all, for the third one (Blaupunkt). In the most difficult part of the route however, the ratings of the second device decrease so much that it falls to the level of the third one. Overall, during this task, interaction with the TomTom personal navigation device was experienced much more positively than interaction with the Blaupunkt navigation device with the Mio Moov falling in between.

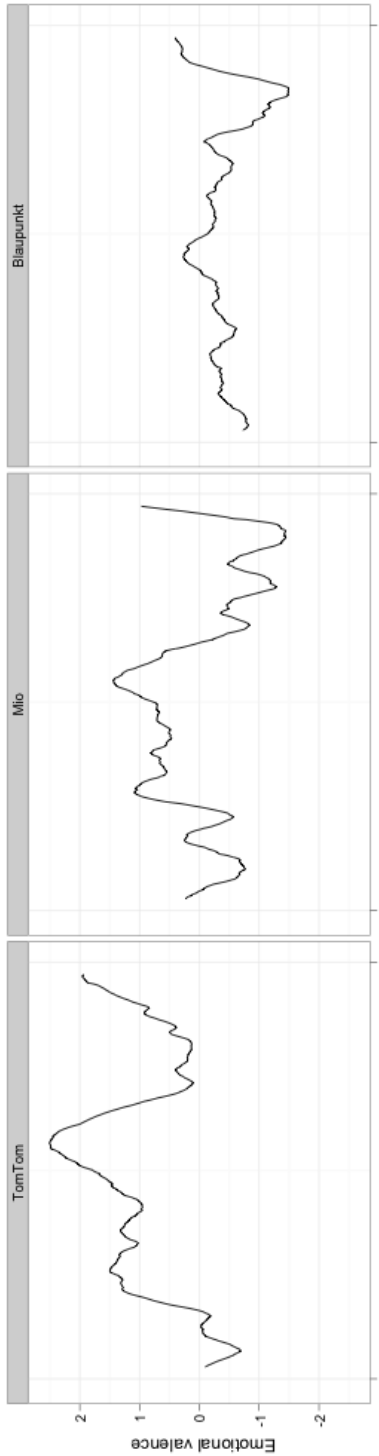


Figure 5.8. *Averaged valence ratings for the first task, broken down by navigation device.*

Experience during the second task, driving back to the university, was generally less differentiated (figure 5.9). Still, the augmented reality navigation device (Blaupunkt TravelPilot) elicited more negative ratings for the first half of the task. In the last part of the interaction, all participants reported somewhat positive emotions as they approach the university, no matter which device they were using. This section of the route was generally easier to follow and the participants would be expected to be familiar with it as they were recruited on the campus (the total time-on-task for the drive back to the university was also much shorter).

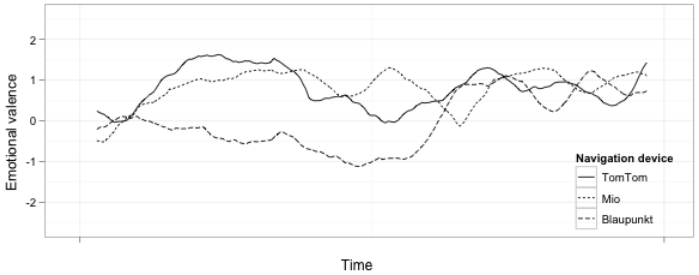


Figure 5.9. *Mean valence ratings for the second task (driving back to the university), broken down by navigation device.*

These moment-to-moment ratings can be compared to the self-reported emotions collected after the drive (figure 5.5). Whereas one of the three devices clearly elicited an inferior user experience during most of the activity, it was not rated more negatively, overall, than the second best device. However, while these results are suggestive, the modest sample size and high variability make any formal modelling of this relationship impossible.

5.2.3. Discussion

The second experiment extended and supported the results from the first experiment and illustrated the ability of self-confrontation to provide meaningful information on emotional experience and on the interaction between a product and its environment of use. A graphical analysis of the moment-to-moment ratings during self-confrontation also provided a detailed picture of the time course of the interaction, revealing differences in self-reported feelings and user experience that were not apparent in questionnaire-based post-test assessment.

An informal comparison between the moment-to-moment experience revealed by self-confrontation and post-test questionnaire data suggests that these two types of measures produced different patterns of differences between products. While this interpretation is

obviously somewhat speculative, this finding can readily be related to the results on the integration of experience obtained in other fields (e.g. Fredrickson & Kahneman, 1993; Redelmeier & Kahneman, 1996).

From this perspective, the mean level of positive or negative experience during an activity does not affect the memory of this activity. Moment-to-moment changes in feelings are not integrated by averaging but by comparing "peak experiences", i.e. the most positive and most negative part of the activity. A product performing badly at some key moment in the interaction (in this case the end of the first task) will therefore be remembered as a product with a poor user experience, even if moment-to-moment ratings suggest that it did in fact also elicit a large amount of positive feelings for most of the time spent interacting with the product. Conversely, a product that did generate higher peak positive experiences and did not perform so badly at its worst will be rated much more positively afterwards, even if it was not that different on average.

5.3. Conclusion

The two experiments reported in this chapter represent the first applications of the approach described in chapter 4. Self-reported ratings of emotional experience collected with the self-confrontation technique were found to be sensitive to momentary changes in feelings and, importantly, to differences between products.

Furthermore, the moment-to-moment ratings in the second experiment revealed dynamic patterns of user experience that were readily interpretable by the researchers and by the participants (in the follow-up interviews). These patterns were not reflected in the traditional user experience questionnaires administered after the test and would be difficult to reconstruct retrospectively without the support of the video.

Finally, the discrepancies between the moment-to-moment data about the user experience and the overall self-reports correspond to important results about the integration of ongoing experience. This finding illustrates the type of research questions that can be addressed using the approach developed in this thesis and the diagnostic value of moment-to-moment measures of emotion for user experience design.

6. The Emotion Slider

The "self-confrontation" procedure described in the previous two chapters is based on the moment-to-moment self-report of their affective state by research participants. In music or advertisement research, this type of self-report is often collected using purpose built devices (e.g. Geringer, Madsen & Gregory, 2004) or a mouse-based graphical user interface (e.g. Schubert, 1999). All these input mechanisms require participants to monitor some form of visual feedback and adjust their response accordingly. The present chapter describes the design of the emotion slider, a device designed to facilitate this process through the use of tangible feedback, and to its empirical evaluation.

6.1. Theoretical background and design of the emotion slider[1]

The starting point of the work presented here is that the collection of moment-to-moment self-report data could benefit from a design perspective. Thus, industrial design is not only used as an object of study or to provide questions and stimuli for applied research but as a purveyor of new approaches or tools for research. In this particular case, the research tool would simplify self-confrontation (see chapter 4) and support the self-report of experience by making the physical interaction with the data collection device as intuitive as possible and reducing the reliance on visual feedback.

The driving question behind this effort became: How can the physical properties and interaction characteristics of a device reflect the feelings of the user? This idea can be related to work in the field of tangible interaction, where the literal correspondence between the interface and the represented information (Blackwell, Fitzmaurice, Holmquist, Ishii & Ullmer, 2007) is a central concept. Recent work in the psychology of emotion around the concept of embodiment provides such a mapping. This body of research suggests that affective responses engage the whole body, not as a consequence but as an integral part of emotion and its representation (Niedenthal, 2007).

1 This section is based in large part on material presented at the *Design Research Society's* 2008 conference and published in its proceedings.

According to this view, even thinking or reflecting upon emotions involves not only symbolic representations but also the expressive, physiological, motivational, and behavioral components of emotion. The basic approach/avoidance tendency (i.e. the tendency to look for positive experience and avoid negative ones) that has been shown to be deeply ingrained in our nervous system as a result of our evolutionary past (Panksepp, 1998) would also be recruited through a process of "motoric reexperiencing". Of course, not every affective process leads to an overt approach or avoidance movement but this embodiment could rely on simulation (Barsalou, 2009), activating the different components of emotion and facilitating subsequent responses congruent with the simulated emotion. Affective self-report would also engage these different systems and the device presented here attempts to capitalize on these powerful forces to provide an effective way to collect data about the affective experience of users.

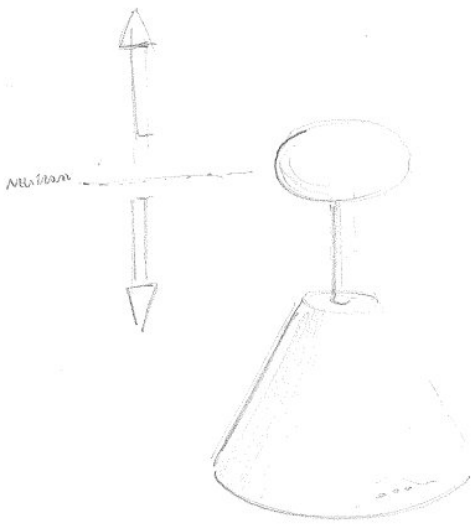


Figure 6.1. *Early sketch for a continuous emotion report device.*

Using the basic approach/avoidance movement as a guiding principle, several designs were considered. Whereas many existing self-report devices are small in size and operated only by the movement of the fingers, our choice went to a relatively large device, resting on a table in front of the participant. Such a device allows bigger amplitude in the movement and easy manipulation by grasping the handle and moving the whole hand. Figure 6.1 presents the first iteration of the selected design.

Figure 6.2. *Photograph of the emotion slider.*

This design then evolved to the current device, dubbed the "emotion slider" (figure 6.2). The vertical movement was replaced with a horizontal sliding movement both for technical reasons and to allow the user to adopt the same hand position while pushing and pulling the handle. The final device is a 40 cm long wooden box with rounded ends. A round shaped handle placed on top can be grasped with one or both hands and pushed or pulled along a rail. The handle and the side of the case are made of bare wood while a metal plate protects the top of the device and hides the springs, sensors and electronic board. The bottom is covered with a leather sheet that can be removed to reveal a stripe of adhesive tape and fix the device solidly to a table.

The further the handle is pushed, the more it resists offering a tangible counterpart to the intensity of the emotion. If left alone, it comes back to the central position, so that a continuing emotion must be reported by constantly pushing or pulling the handle to maintain it in position.

6.2. Empirical evaluation

Before using it to measure product experience with self-confrontation, the emotion slider was subjected to a series of experiments to test empirically the hypotheses underlying the device and assess its usefulness for research on affective experience. Several studies were thus conducted to find out whether the ideas and hypotheses regarding the interaction with the slider and its properties are warranted. Specifically, the main hypothesis is that approach-avoidance tendencies are activated through the evaluation of affective stimuli, as theories on

the embodiment of emotion would suggest, and that these tendencies would make specific movements easier or more intuitive.

As noted in chapter 4, there is a growing body of empirical research on affect-movement compatibility, showing that the processing or evaluation of affectively valenced stimuli facilitates specific movements and inhibits others. Chen and Bargh (1999), inspired by an early experiment by Solarz (1960), obtained shorter response times (taken as an indicator of congruence) from participants asked to evaluate words by pushing a lever to report a negative evaluation and pulling it to report a positive evaluation than from participants assigned to a reversed set of instructions (pulling the lever for negative words and pushing it for positive words). Chen and Bargh interpret this effect as evidence for the existence of an adaptive back-up system, automatically promoting approach (arm flexion, e.g. to pull something toward oneself) and withdrawal (arm extension, e.g. to push an aversive stimulus away).

Following this paper, a series of publications on the topic appeared, focusing on the automacity of the effect (Duckworth, Bargh, Garcia & Chaiken, 2002; Rotteveel & Phaf, 2004) and on the specificity of approach-avoidance effects to certain emotions, such as fear or anger (Alexopoulos & Ric, 2007; Marsh, Ambady & Kleck, 2005).

More recently, as the present research was underway, a number of results have called into question the idea of a direct mapping between valence and arm movement, stressing the flexibility of approach or avoidance depending on the consequences of the movement or the frame of reference induced by the procedure but still documenting many examples of affect-movement compatibility effects (Bamford & Ward, 2008; Eder & Rothermund, 2008; Seibt, Neumann, Nusinson & Stark, 2008; Van Dantzig, Pecher & Zwaan, 2008).

In light of this literature, an investigation of the consequences of affect-movement compatibility effects on moment-to-moment self-report of emotion seems warranted. It should also be noted that almost all of the results described above are based on the evaluation or the detection of single words or facial expressions and the accuracy of the ratings is not usually a focus of the research. Beyond testing the ideas underlying the design of the device, experiments with the emotion slider can also provide some information on the impact of approach-avoidance tendencies on the measurement process and whether this should be a concern for researchers collecting affective self-report data with similar and not-so-similar devices.

The general approach adopted to test the emotion slider and the ideas underlying its design is modeled after the literature on movement-affect compatibility. In each experiment, a condition in which the slider is used in the intended way, hypothesized to be congruent with the affective response to be reported, is contrasted with a control condition in which the slider is not used in the intended

way, typically inverting the direction of the movement asked from the participants. The primary outcome is a comparison of the response times in each condition. In this context, a quicker response time is not taken to be desirable in itself but is used as an index of congruence. If the approach system is activated by a stimulus evaluation, reporting this evaluation with an approach movement should be quicker and the use of the emotion slider the way it was designed should be facilitated.

Another outcome that will be examined is the accuracy of the evaluations. Published accounts of research on affect-movement compatibility typically mention errors in passing, mostly to rule out a speed-accuracy trade-off by the participants. From a practical point of view however, accuracy is of great importance. If a particular response modality turned out to improve or reduce self-report accuracy, this would be a major concern for researchers collecting such data.

While the emotion slider and other similar devices were obviously designed to be used continuously with dynamic stimuli, the experiments presented here all use static stimuli, namely photographs. The reason for this choice is twofold: well-known, standard stimuli are readily available in this form and still pictures allow for an unambiguous definition of response time as the time elapsed since the onset of the picture. Films would have been even closer to the intended use of the emotion slider and several sets of clips selected for their emotional content can be found in the literature, but it can be difficult to attribute affective responses to specific events or time points in the movie and therefore to measure how quick the response was.

However, even static pictures are vastly more complex than the stimuli used in previous research. Testing the emotion slider by collecting affective ratings of photographs therefore seems a useful way to bridge the literature on affect-movement compatibility and research on the measurement of emotion, providing some insights into the relevance of approach/avoidance tendencies in situations broadly similar to product experience research.

6.2.1. Experiment 1[2]

Experiment 1 was the first test of the emotion slider, focusing on the correlation between slider movement and normative valence ratings of the stimuli used[3].

2 Data from this experiment were used in a paper presented at the *Design Research Society's* 2008 conference and published in its proceedings.

3 I am grateful to Max Braams, Maarten Langbroek and Jorn Ouborg for their help in setting up and carrying out this experiment.

6.2.1.1. Stimuli

The stimuli used in this experiment were photographs of life scenes extracted from the International affective picture system or IAPS (Bradley & Lang, 2007; Lang, Bradley & Cuthbert, 2008). These pictures are widely used in affective science to elicit emotions. They are selected for their affective content and come with normative ratings on three dimensions: valence, arousal and dominance.

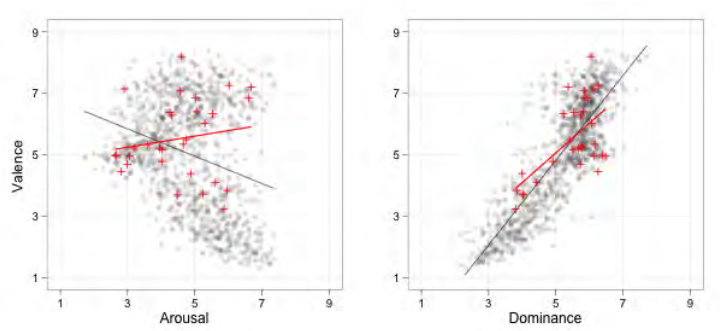


Figure 6.3. *Mean normative ratings for IAPS pictures: valence (vertical axis), arousal (left) and dominance (right)*[4]. *Grey dots represent the whole set, red "+" represent pictures selected for experiment 1*[5].

The slides used in experiment 1 were picked randomly from the IAPS, taking several steps to ensure that the selected pictures represent a broad sample of affective material varying by variance and arousal. First, gender-specific stimuli (i.e. pictures eliciting widely different ratings from men and women) were removed from the set, which excluded many erotic pictures. Gruesome accident or injury pictures were also excluded for two reasons. Firstly, the type of affective response elicited by such picture does not seem very relevant for a design-oriented research project. Secondly, even though IAPS pictures are not very different from the material typical encountered on TV, exposing participants to even mildly disturbing stimuli would not be justified given the purpose of this experiment.

After filtering the picture set, the slides were ordered by increasing valence based on the IAPS norm and divided in five groups, randomly

4 The IAPS technical manual includes normative ratings collected with two different version of the SAM dominance scale (Lang, Bradley & Cuthbert, 2008). When both were available for a given picture, only the ratings from the older scale were used for the graph.

5 The trend line is a linear regression line constructed with the *geom_smooth(method="lm")* function in R's *ggplot2* package (Wickham, 2009).

picking three photographs in each group of pictures. The same procedure was then followed using the arousal ratings, yielding a total of 30 pictures. Using such a relatively large number of stimuli is typical in IAPS research and has several advantages. Multiple pictures afford several trials in each condition, compensating for the noisiness of low reliability measures (e.g. response time) and potential individual differences in response to individual pictures. It also ensures that the picture set includes a variety of content and samples broadly from the affective dimensions, which is necessary to obtain meaningful correlations between these affective dimensions and other variables.

The pictures selected for this experiment have the following codes in the IAPS: 1026 – snake, 1110 – snake, 1440 – seal, 1616 – bird, 1731 – lion, 2092 – clowns, 2191 – farmer, 2351 – nursing baby, 2370 – three men, 2495 – man, 2682 – police, 2690 – terrorist, 4598 – couple, 4613 – condom, 4624 – couple, 4680 – erotic couple, 4695 –erotic couple, 6930 – missiles, 7030 – iron, 7034 – hammer, 7035 – mug, 7182 – checkerboard, 7185 – abstract art, 7224 – file cabinets, 7450 – cheeseburger, 8117 – hockey, 8490 – roller coaster, 8600 – mascot, 9160 – soldier, 9270 – toxic waste. The normative ratings for these pictures in the valence-arousal-dominance space are shown in figure 6.3, together with the rest of IAPS stimuli.

6.2.1.2. Participants and procedure

Participants (N = 39, 23 men and 16 women) were students at Delft University of Technology who volunteered to participate. Since the data from two participants were lost due to a technical problem; the following discussion is based on an effective sample size of 37 participants.

The participants were first asked to read and sign an informed consent form and to fill in the Dutch version of the PANAS, with "current mood" instructions (Peeters, Ponds & Vermeeren, 1996). They were then seated in front of a laptop computer with the emotion slider attached to the table in front of the computer. The computer was running a purpose-built VB.NET software. The procedure was explained by means of an on-screen introduction, including three example stimuli (IAPS codes 3300, 5833, and 7010) to give participants an impression of the range of pictures they could expect. About half of the participants (N = 16 from 37) were invited to report positive feelings by pushing on the handle and conversely to report negative feelings by pulling it. The rest of the participants received the opposite set of instructions (push to report negative feelings and pull to report positive feelings). After going through all the pictures in a random order, the participants were asked to fill in a brief *ad hoc* questionnaire about the device.

6.2.1.3. Results

The first type of data examined in experiment 1 is the movement of the slider itself. For each trial, the software controlling the device recorded the amplitude of the movement, defined as the distance between the rest position of the handle and the farthest points reached by the handle while the picture was present on screen. The resolution of the device allows a measurement of this distance on a scale from -11 to +11. If no movement was recorded (i.e. the handle remained in the rest position) a score of "0" was entered. Individual distances were averaged across participants to provide a mean distance from the center for each picture in the set. These mean distances were compared to the normative valence ratings provided with the IAPS (figure 6.4), r = .90 (95% modified percentile bootstrap confidence interval: [.81, .96])[6]. This correlation seems slightly lower for the group pushing for positive pictures (r = .84, 95% CI: [.68, .92]) than for the group pushing for negative pictures (r = .93, 95% CI: [.87, .97]) but there is a large overlap between the two confidence intervals.

6 All confidence intervals for Bravais-Pearson product-moment correlation coefficients in this chapter are based on the modified percentile bootstrap method developed by Wilcox (1996), see Wilcox (2003), pp. 216-218. They were computed with the *pcorb* R function by Rand Wilcox (see R's *WRS* package and Wilox, 2005, p. 403).

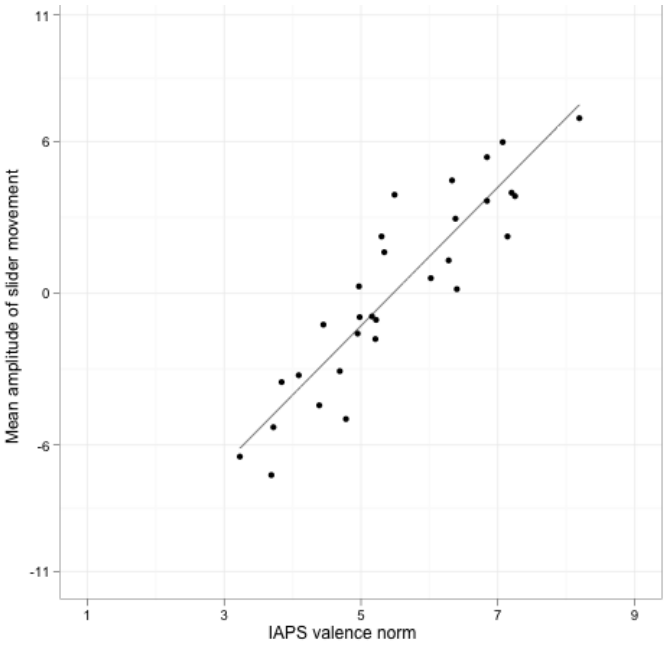


Figure 6.4. *Scatterplot of mean amplitude of slider movement for each picture against normative IAPS valence score.*

The high correlation between slider movement and the normative valence ratings can be compared with the correlations between slider movements and the other dimensions of affect documented in the IAPS norm (table 6.1). This comparison shows the valence as measured by the emotion slider has a higher correlation with the valence measured by the paper-and-pencil SAM than with any other dimension measured by the same method. Slider movement amplitudes also exhibit various levels of association with IAPS normative ratings of arousal and dominance but the pattern of these correlations corresponds closely to the magnitude of the associations between these two dimensions and the normative valence ratings themselves. The relatively large correlation between valence and dominance is not only apparent in the ratings of the stimuli used in this experiment (r = .63) but also in the whole set of over 1000 slides in the IAPS (r = .84, see also figure 6.3) and can therefore be interpreted as a property of the pictures themselves rather than a lack of specificity of the slider. Similar patterns of correlation between valence, arousal, and dominance have in fact been observed in other situations, such as ratings of emotion-eliciting situations collected with multi-item verbal scales (Russell & Mehrabian, 1977).

Table 6.1. *Correlations between slider movement and IAPS ratings.*

	amplitude	valence	arousal	dominance
slider amplitude	1	.90	.16	.64
IAPS valence		1	.17	.63
IAPS arousal			1	-.44
IAPS dominance				1

Another important aspect of the ratings collected with the slider is their accuracy. Unfortunately, what should count as an erroneous trial is not obvious when considering affective self-reports or evaluations. Firstly, since the pictures included in this experiment were selected to span a wide area of the affective space, some of them are only mildly positive or negative or have a rather neutral valence. Consequently, a non-response can represent both a slip of attention or a valid "neutral" response. Secondly, and more importantly, current theories of emotions stress that affective responses are shaped by one's appraisal of the environment, its dangers and opportunities, relative to one's goals, beliefs and life experience. Some variability is therefore expected, even if a relatively passive laboratory situation and the innocuousness of the pictures can be expected to limit the personal involvement. As an example, picture 9001 represents a graveyard in winter and is typically rated as strongly negative but it is conceivable that focusing on the aesthetic quality of the picture or failing to recognize its symbolic charge might prompt someone to sincerely rate it as positive or neutral. The important point is that while IAPS pictures were selected to elicit specific affective ratings, this does not necessary mean that every atypical self-report is a mistake. Subsequent experiments employed two strategies to deal with these difficulties but for experiment 1, differences in the number of non-responses between the two conditions were tested as a proxy for actual mistakes, keeping in mind that this count is at best a noisy indicator of incorrect trials, since many non-responses actually reflect a genuine neutral rating.

The last type of data examined in this experiment is the time necessary for the participant to initiate a movement of the slider. Published studies on approach-avoidance facilitation always use similar experimental designs, with multiple trials in each cell of the designs and analysis with simple univariate ANOVAs on mean cell response times (e.g. Alexopoulos & Ric, 2007; Bamford & Ward, 2008; Chen & Bargh, 1999; Duckworth, Bargh, Garcia & Chaiken, 2002; Eder & Rothermund, 2008; Marsh, Ambady & Kleck, 2005;

Rotteveel & Phaf, 2004; Seibt, Neumann, Nusinson & Stark, 2008; Van Dantzig, Pecher & Zwaan, 2008). In most situations, this type of analysis leads to an underestimation of the type I error rate and it has long been recognized as incorrect in other subfields of psychology (Clark, 1973). However, in the series of experiments reported in this chapter, the key manipulation is a between-subject factor and the exact same pictures are used in each condition. In this particular situation, a regular univariate ANOVA or t-test on the participants' mean response times is appropriate (Raaijmakers, Schrijnemakers & Gremmen, 1999). Other data analysis techniques (in particular mixed-effects modeling; Baayen, Davidson & Bates, 2008) can provide more flexibility and power but the simpler approach is a "minimally sufficient analysis" as recommended by Wilkinson and the Task Force on Statistical Inference (1999).

Response time was defined as the time between the onset of the picture and the moment a movement of the handle was registered by the slider. Trials during which no movement was recorded were treated as missing data. One outlier (*RT* = 31 ms) was also removed before all analyses. The remaining response times were averaged across trials to yield a mean response time for each participant. These average response times were very similar in both group of participants, with a mean response time of 2860 ms (*SD* = 725 ms) for the group asked to push the handle for positive pictures and 2855 ms (*SD* = 441 ms) for the group asked to push the handle for negative pictures (figure 6.5). The observed sample difference is very small and a t-test (with Welch correction for unequal variances) also indicates that there is no evidence for a difference in average response time, $t(23.249) = -.03$, $p = .98$, Cohen's $d = -.01$ (95% confidence interval for the difference: [-430 ms, 416 ms])[7].

7 T-tests for differences in response times were performed with the *t.test* function in R's *stat* package. By default, this function uses Welch's t-test with the Welch-Satterthwaite's correction to the degrees of freedom to account for (potential) differences in variance between the two groups. The results can therefore differ from those that would be obtained with software (e.g. SPSS/PASW) using Student's t-test and a pooled variance estimate. For experiment 1, the correction is rather large because the two sample standard deviations are far from equal. The resulting confidence interval is therefore noticeably wider than an uncorrected confidence interval (in this case [-396 ms, 385 ms]). The difference is not as large in other experiments.

Regarding effect size, Cohen (1977, p. 20) does not specify the standard deviation to use to compute standardized mean differences, as it is supposed to be equal in both populations. As is common, standardized effect sizes in this chapter were computed used a pooled variance estimate (Thompson, B., 2007). Obviously, the large variance difference between groups in experiment 1 does not only impact the test results but also this standardized effect size.

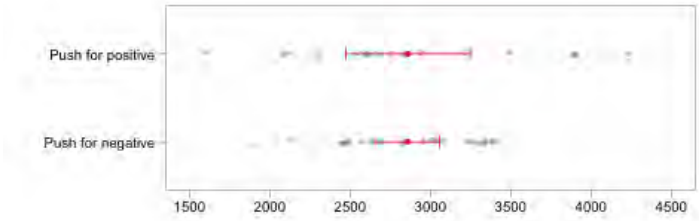


Figure 6.5. *Response times (in ms) for experiment 1. Each grey dot represents the mean response time for a single participant. The red dots and error bars represent the point estimate and 95% CI for the mean response time in each group (participants pushing the handle for positive stimuli vs. participants pushing the handle for negative stimuli).*

6.2.1.4. Discussion

The amplitude of the movement of slider handle is strongly correlated with the valence scores in the IAPS norm, despite the lack of visual feedback and the fact that participants were not instructed to make any distinction beyond a binary positive/negative classification. This finding suggests that the emotion slider provides an intuitive representation of emotional valence and that visual feedback is not necessary for participants to consistently express their feelings with it.

However, comparisons between the two conditions do not reveal any clear congruency effect. The variability of the response times is quite high and the confidence interval of the difference in response time is very broad. The data presented here is therefore compatible with anything from a typical congruency effect (differences reported in similar studies in the literature are all under 300 ms) to a strong effect in either direction or no difference at all. This high variability might have resulted from the lack of emphasis on speed in the instructions and the choice of pictures, which included neutral stimuli, unlike most published experiments about affect-behavior congruence.

6.2.2. Experiment 2[8]

Another experiment was conducted to further investigate congruency effects between valence and movement direction and to alleviate the issues identified in the discussion of the results of the first

8 I am very grateful to Remon de Wijngaert for his great help in planning and conducting experiment 2 and 3 with the emotion slider. Data from experiment 2 served as the basis for a paper presented at the *Affective Computing and Intelligent Interaction 2009* conference and published in its proceedings.

experiment. Two aspects of the procedure were changed to improve power and to try to replicate published congruency effects: speed and choice of pictures. Speed was increased by emphasizing quick response in the instructions and reducing the length of time each picture was displayed. As the (within-group) variance in response time distributions is well known to increase with the mean (Wagenmakers & Brown, 2007), reducing the average response time guarantees more power to detect potential between-group differences, as long as the difference itself is stable. Additionally, the set of stimuli was revised to avoid including neutral valence/low arousal pictures.

6.2.2.1. Stimuli

Another set of IAPS pictures was prepared for this experiment. These slides were selected in two groups: 10 positive pictures (1440 – seal, 1441 – polar bears, 1463 – kittens, 1710 – puppies, 2070 – babies, 2388 – kids, 5760 – nature, 5833 – beach, 7330 – ice creams, 8380 – athletes) with an average normative valence rating between 7.44 and 8.34 and 10 negative pictures (2683 – war, 2703 – sad children, 2900 – crying boy, 3280 – dental exam, 7380 – roach on pizza, 9001 – cemetery, 9041 – scared child, 9290 – garbage, 9300 – dirty, 9902 – car accident) with an average normative valence rating between 1.91 and 3.72. The normative ratings for these pictures in the valence-arousal-dominance space are shown in figure 6.6.

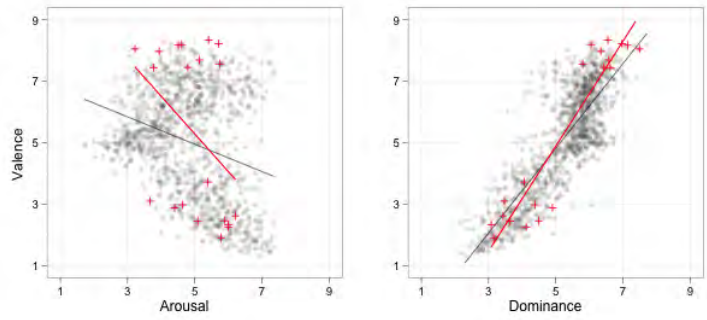


Figure 6.6. *Mean normative ratings for IAPS pictures: valence (vertical axis), arousal (left) and dominance (right). Grey dots represent the whole set, red "+" represent pictures selected for experiment 2, 3, and 4.*

6.2.2.2. Participants and procedure

Participants (N = 51, 36 men and 15 women) were students at Delft University of Technology who volunteered to participate. The procedure was similar to the one used in experiment 1, except for some slight change in the instruction and stimulus presentation:

the participants were invited to report their evaluation as quickly as possible and the pictures were displayed for only 2s to encourage a quick response. As in experiment 1, half of the participants (N = 26 from 51) were asked to push the slider for negative pictures and the rest was asked to push the handle for positive pictures. After going through the whole set of pictures, participants were also asked to review their responses one by one and indicate which one were in fact errors.

6.2.2.3. Results

As in experiment 1, the amplitude of the handle movement was recorded and averaged over pictures. The mean movement amplitude correlates highly to the normative IAPS ratings in all conditions: r = .98 (95% CI: [.96, .99]) for participants asked to push for negative pictures and r = .99 (95% CI: [.98, 1.00]) for participants asked to push for positive pictures.

Response times for all correct trials were averaged across trials and the mean per-participant response times were used to compare both conditions. All atypical trials were removed from the data set prior to these analyses. Three types of trials were thus removed: trial with no response before the offset of the picture, responses subsequently reported as erroneous by the participants and unexpected responses (i.e. positive evaluation for a picture with a negative valence score in the IAPS norm and vice versa). As shown on figure 6.7, the participants asked to push for negative pictures were slower (M = 907 ms, SD = 130 ms) than the participants pushing for positive pictures (M = 833 ms, SD = 111 ms). The difference is significant at the conventional 5% level, $t(48.36) = 2.18$, p = .03, Cohen's d = .62 (95% CI for the difference: [6 ms, 142 ms]).

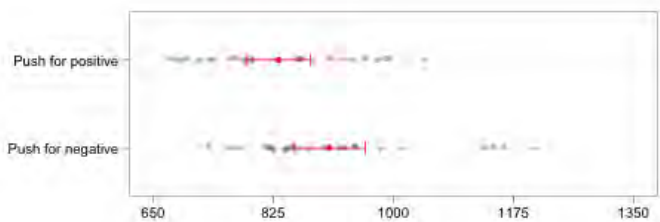


Figure 6.7. *Response times (in ms) for experiment 2. Each grey dot represents the mean response time for a single participant. The red dots and error bars represent the point estimate and 95% CI for the mean response time in each group (participants pushing the handle for positive stimuli vs. participants pushing the handle for negative stimuli).*

The self-reported error count was investigated with a logistic regression model, using "push for positive" as the reference group. A test of the deviance difference between the null model and a model using the direction as predictor is barely significant: χ^2 (1) = 3.87, p = .05 (95% CI for the odds of the difference: [1.0, 3.8]).

6.2.2.4. Discussion

This experiment revealed a clear valence-related facilitation effect, on a magnitude similar to the effects reported in the psychological literature. However, the direction of the effect did not conform to the prediction and the congruent instruction set was "push for positive", prompting two further experiments detailed below.

Furthermore, the error rate seems somewhat lower in the congruent condition, ruling out a speed-accuracy trade-off and suggesting that affect-movement congruency might have some impact on measurement based on the emotion slider and similar devices. Still, the difference, if any, is quite small and the error rate was very low in all conditions (between 94% and 96% accuracy).

Correlations between the amplitude of the movement and normative valence ratings of the pictures were very high in both conditions. They were also higher than in the first experiment but this is to be expected with a stimulus set including only pictures with "extreme" (positive or negative) valence. Consequently, the correlations with normative ratings observed in this experiment cannot be interpreted as evidence for the validity of slider measures over the whole range of affective pictures in the IAPS.

6.2.3. Experiment 3

Experiment 2 showed that a clear congruency effect with a small but noticeable effect on the accuracy of the evaluation is present when using the emotion slider to rate pictures. This effect was however not in the same direction as the bulk of published results in the literature at the time and another experiment was set up to attempt to recover the original effect and help interpret the inverted effect of experiment 2. The original hypothesis was based on a link between arm extension and avoidance tendencies (pushing dangerous objects away) or arm flexion and approach tendencies (pulling pleasurable things towards oneself). Arguably, the mapping between arm flexion and extension on the one hand and approach and avoidance on the other hand is not totally unambiguous. In the experimental situation described above, pushing on the handle could also be interpreted as an approach movement, literally bringing the participant closer to the stimulus. Conversely, pulling could be interpreted as an avoidance movement,

getting away from the stimulus.

To remove this ambiguity, the procedure was changed to add visual feedback in the form of a variable picture size controlled by the movement of the slider's handle. As the participants pushed on the slider, the pictures would shrink, emphasizing the movement *away* from the body. Conversely, pulling on the slider would make the picture grow, as if the participants were pulling it towards them. Similar manipulations have been used by Bamford and Ward (2008), Van Dantzig, Zeelenberg, and Pecher (2009), or Markman and Brendl (2005).

6.2.3.1. Participants and procedure

Participants (N = 43, 31 men and 12 women) were students at Delft University of Technology who volunteered to participate. The procedure was identical to experiment 2. Half of the participants (N = 22 from 43) were asked to push the handle for negative stimuli, while the rest pushed for positive stimuli.

To reduce the ambiguity in the movement elicited from the participants, a new form of visual feedback was introduced: a forward movement of the handle (i.e. away from the participant's body) made the picture shrink, while a backward movement caused the picture to grow. The visual feedback was constant across conditions so that pushing on the slider would always result in a shrinking picture, no matter the instructions (pushing for positive vs. pushing for negative).

6.2.3.2. Results

Trials with response times less than 200 ms (4 out of 880) were deleted from the data set before conducting the analyses.

As in other experiments, the amplitude of the handle movement was recorded and averaged over pictures. The mean movement amplitude correlates highly to the normative IAPS ratings in all conditions: r = .98 (95% CI: [.96, .99]) for participants asked to push for negative pictures and r = .99 (95% CI: [.98, 1.00]) for participants asked to push for positive pictures.

Response times for all correct trials were averaged over trials and the mean per-participant response times were used to compare both conditions. Participants in both conditions responded at virtually the same speed: M = 966 ms (SD = 180 ms) for participants pushing for negative pictures and M = 934 ms (SD = 124 ms) for participants pushing for positive pictures (figure 6.8), $t(37.442) = 0.69$, $p = 0.5$, Cohen's d = .21 (95 % CI of the difference: [-62ms, 127ms]).

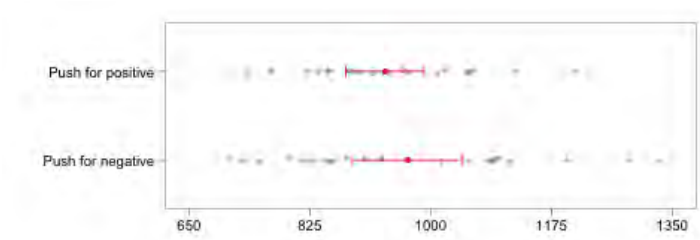


Figure 6.8. *Response times (in ms) for experiment 3. Each grey dot represents the mean response time for a single participant. The red dots and error bars represent the point estimate and 95% CI for the mean response time in each group (participants pushing the handle for positive stimuli vs. participants pushing the handle for negative stimuli).*

The difference in the number of self-reported errors is also small and not significant (95% CI for the odds of the difference: [0.8, 3.6]).

6.2.3.3. Discussion

While a non-significant result does not provide much evidence of equivalence (Cohen, 1999; Loftus, 1996; Tukey, 1991), in this case the observed difference is very small and the response time distribution for both groups almost completely overlap. These results strongly suggest that no congruence effect is present in this data and that the visual feedback does have an effect on approach-avoidance tendencies, essentially cancelling the effect obtained in experiment 2.

6.2.4. Experiment 4[9]

The results of experiment 3 suggested that congruence effects are more malleable than suggested by the earlier part of the literature but did not fully elucidate the reasons for the direction of the effect measured in experiment 2. Another interpretation of this effect was often mentioned during informal conversations with colleagues and visiting scientists is the possibility that pushing was associated with positive valence through the activation of an UP-DOWN image schema. This hypothesis received some support from the literature (Meier & Robinson, 2004) and participants also spontaneously speak of the movement of the slider as if it was along a vertical dimension during pilot studies and debriefing interviews.

Experiment 4 was conducted to further investigate this hypothesis and deconfuse the context-bound approach movement from the

9 I am grateful to Ahmet Bektes for his help in setting up and conducting this experiment.

mapping with the vertical dimension. To achieve this, the experimental situation was altered to put the slider beside the screen, ensuring that moving the slider's handle would not result in any change of the participant's position relative to the stimuli. If the congruence between the "push" movement and positive evaluation is indeed driven by the activation of an UP-DOWN image schema, the effect should remain as strong as in the previous situation (experiment 2), when the slider was placed between the screen and the participant.

6.2.4.1. Participants and procedure

Participants in this experiment (N = 50, 21 women and 29 men) were master-level students in Industrial Design Engineering at Delft University of Technology who volunteered for participation. After giving consent, the participants were asked to fill in the I-PANAS-SF (Thompson, 2007) and TIPI scales (Gosling, Rentfrow & Swann, 2003). The procedure was identical to the one used in experiment 2, save for the fact that the screen was a laptop screen laying horizontally on the table in front the participant. The slider was attached to the table, to the right of the screen. For this reason, the participants who reported using the computer mouse with the left hand and requested the device to be placed on the other side of the screen were excluded from the analysis. Participants who reported having seen the pictures used in the experiment before (presumably in other experiments running at the same time) were also removed from the data set, yielding a final sample size of 39 participants.

6.2.4.2. Results

As in the other experiments, the amplitude of the handle movement was recorded and averaged over pictures. The mean movement amplitude correlates highly to the normative IAPS ratings in all conditions: r = .98 (95% CI: [.97, 1.00]) for participants asked to push for negative pictures and r = .99 (95% CI: [.98, 1.00]) for participants asked to push for positive pictures.

Response times for all correct trials were averaged over trials and the mean per-participant response times were used to compare both conditions. The participants asked to push for negative pictures were apparently somewhat quicker (M = 870 ms, SD = 141 ms) than the participants pushing for positive pictures (M = 936 ms, SD = 164 ms) but the difference was not significant (figure 6.9), $t(31.593) = -1.31$, p = .20, Cohen's d = -.37 (95% CI of the difference: [-167 ms, 36 ms]). This experiment therefore failed to find a clear congruency effect in either direction.

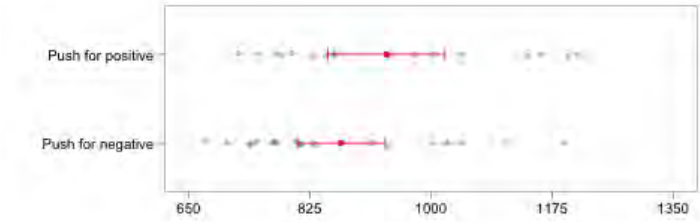


Figure 6.9. *Response times (in ms) for experiment 4. Each grey dot represents the mean response time for a single participant. The red dots and error bars represent the point estimate and 95% CI for the mean response time in each group (participants pushing the handle for positive stimuli vs. participants pushing the handle for negative stimuli).*

While the result of the statistical test indicates that the presence and direction of any potential effect is uncertain, effect sizes can still be used to compare the outcome of this experiment with previous ones. Interestingly, the confidence intervals for the response time difference (i.e. the unstandardized effect size; Baguley, 2009) suggest that, whatever its direction, the difference between the groups when the slider is place beside the screen (experiment 4) is less than[10] than the difference obtained with a slider in front of the screen (experiment 2).

The number of self-reported errors was similar in both conditions (95% CI for the odds of the difference: [0.5, 1.6]).

6.2.4.3. Discussion

Even if the evidence for a congruency effect in this experiment is weak at best, the data hints[11] towards a small speed advantage for the "push for negative" condition. This difference is however not significant at the 5% threshold, which means that the direction of the effect cannot be formally established at this error rate. Accordingly, the confidence interval of the difference in means includes 0, even if it also suggests that the difference is more likely to be positive than negative. However, even if it were negative this difference would be very small indeed, an order of magnitude smaller than the effects typically reported in the psychology literature.

Furthermore, the difference in mean response times between the two conditions is significantly lower than that obtained in experiment 2, clearly contradicting the hypothesis that the congruency effect

10 "Less than" is to be understood numerically, i.e. it is either a relatively large negative difference (i.e. a difference in the opposite direction) or a small positive difference but in any case not a large difference in the same direction than in experiment 2.

11 To use Tukey's terminology (see Abelson, 1995).

would be sustained or reinforced if "up" and "avoidance" were deconfounded. The response time data discussed above are clearly incompatible with the interpretation of the slider's operation in terms of vertical movement spontaneously offered by colleagues and participants.

6.3. General discussion

While some of the results are somewhat unclear, this series of experiments allows a number of conclusions on affect-behavior congruence and its influence on the measurement of affect with the emotion slider. These conclusions will be discussed by examining three sets of results from the studies described above, namely correlations with the stimuli normative ratings, response times and accuracy.

The first set of results pertains to the amplitude of the movement exerted on the handle of the slider by the participants. In all the experiments conducted with the emotion slider, this amplitude was highly correlated with the normative SAM ratings for the IAPS pictures (table 6.1). These correlations are somewhat less informative for the three experiments using only relatively strong positive or negative pictures but the high correlation observed in experiment 1 supports the validity of the data collected with the emotion slider as a measure of valence. This is especially interesting because the participants were instructed to report any positive or negative feeling they might experience but not to make gradual valence ratings. The linear relationship between slider movements and ratings collected with a more traditional paper-and-pencil instrument therefore suggest that the shape and physical characteristics of the slider offered a tangible counterpart to the level of valence and was intuitively used to make finer distinctions, at least by some participants.

The second set of results pertains to the response time of the participants when registering their ratings. In this context, a quicker response time is not so much a goal in and of itself than a sign of congruence between the response and the stimuli and a way to assess the effect of the embodiment of emotion on the self-report process. Of all the experiments presented here, the only one demonstrating a clear congruence effect is experiment 2, but this effect (to wit, pushing on the slider handle is congruent with positive affect and pulling is congruent with negative affect) is in a direction opposite to the initial hypothesis (based on the early literature on approach-avoidance effects).

While the two follow-up experiments did not produce a clear congruence effect in the other direction, they did shed some light on the reasons for this mismatch. Considered together, they establish that congruence effects are much more malleable and contingent

that initially thought, as they can be cancelled or inverted by factors such as visual feedback (experiment 3) or the relative position of the participants, slider, and stimuli (experiment 4).

Other results that appeared in the literature while this research was under way can also help interpret these data. Seibt et al.'s (2008) third experiment shows that an affect-motor compatibility effect can be inverted by inducing another "frame of reference" in the instructions. Eder and Rothermund (2008) also measured changes in the direction of the congruence effect depending on instructions, obtaining for example an inversion of the effect when describing the same movement (pushing on a joystick) as "upwards" instead of "away". In all experiments, the instructions were delivered to the participant on screen to ensure consistency and carefully avoided any implication regarding the frame of reference (i.e. participants were asked to "push", not to "push way" or "push toward the screen"). Most other published reports do not clarify exactly what set of instructions were used but authors tend to describe the movement as "pushing away from" or "pulling towards" oneself. If this is also how it was communicated to research participants, it might account for the discrepancies between the results of experiment 2 and earlier studies.

Bamford and Ward (2008) and Van Dantzig et al. (2008) describe experiments demonstrating the impact of repeated visual feedback or "action effects" following a response on the interpretation of a movement as approach or avoidance. This effect certainly accounts for the difference between experiments 2 and 3. It should be noted however that the manipulation used in experiment 3 did not invert the direction of effect but merely cancelled it.

Combined with my own data, these results suggest that the most likely explanation of the results of experiment 2 remains an approach-avoidance effect and that in the absence of conflicting cues (such as visual feedback or specific instructions), the « push » movement is perceived as an « approach » movement toward the screen and the stimulus.

The last set of results pertains to the accuracy of the evaluations. In all cases, accuracy was very good across the board with very few trials self-reported as errors. Nonetheless, in experiment 2 the affect-movement mapping that was most congruent based on the response time data also produced significantly more accurate ratings. The confidence interval of the difference suggests that the number of errors could range between being almost equal to three times bigger in the incongruent condition.

6.4. Conclusion

This chapter described the development of the emotion slider, a device designed to use principles from tangible design and theories about the embodiment of emotion to make moment-to-moment self-report of emotion as intuitive as possible. A series of experiments conducted with the emotion slider compared response times in different conditions to test the ideas behind the design.

These experiments identified an association between specific movements and emotions elicited by pictures but not in the predicted direction. Further experiments also revealed that this congruency effect is in fact very sensitive to contextual factors such as action effects, instructions and physical setting. In any case, the impact on error rates and accuracies remains limited.

If a similar device must be used to measure emotions, the most intuitive mapping in these experiments, namely "pushing" for positive valence and "pulling" for negative valence, with the slider placed between the participants and the screen, would nevertheless seem to be recommended.

7. On Reliability

Reproducibility is a key aspect of any measurement. For a measure to be said to quantify some characteristic of designs or products, it should be possible to obtain similar measures in a reasonably broad range of situations involving these products.

The magnitude of the difference between several replications of the same measurement depends on the amount of error in each individual measurement. The more error there is in the measurement process, the more variation can be expected in successive measures of the same product. Two types of measurement error can be distinguished: systematic and random error. Systematic error affects all products equally (constant error) or perhaps only a group of products or participants (bias). In psychometrics, these types of errors are (a small part of) validity issues whereas reliability quantifies random measurement error and reproducibility. Reliability is therefore related to the notion of precision in physical measurement and efficiency in statistics and conditions the quality and usefulness of all measures.

7.1. Reliability and measurement error

Psychometrics primarily uses two notions to describe the quality of psychological measures: validity and reliability. Validity refers to the meaning and correct interpretation of measures, whether they actually quantify the construct they are supposed to measure, potential bias in the measurement process, etc. Some validity issues are therefore related to the notion of accuracy in physical measurement. However, even a perfectly accurate or valid measurement process is likely to produce slightly different values when repeated several times. In psychometrics, this variability is discussed under the name of "reliability". This terminology departs from the usual sense of the word "reliability". In the common acceptation of the term, a test or method is said to be unreliable because it yields erroneous results. This meaning of the word "reliable" is more akin to the psychometric notion of validity. In fact, as noted by Feldt & Brennan (1989), from a psychometric point of view, a medical test can be very reliable even if it is often wrong, as long as it consistently gives the same diagnostic (true or false) for a given patient.

In psychometrics, reliability is therefore strongly related to (random) measurement error and what is called precision in

physical measurement. Reliability and measurement error limit the reproducibility of psychological measures. A reliable measurement process will produce consistent results across repetitions and allow researchers to confidently generalize their findings to a broader range of situations. In this chapter, reliability will be formalized in the context of classical test theory[1] before considering some issues facing researchers willing to apply it to design-related measurement.

Classical test theory makes some assumptions to be able to derive information about unobservable quantities (e.g. measurement error) from test data. It subsumes several additive "true score" models expressing observed scores in psychological tests as a sum of a true score and a random component:

$X_1 = T + E$ where X_1 is an observed score, T is the true score and E is assumed to be pure random error.

An individual's true score is defined as the (hypothetical) sum of scores on all potential measures (items or tests) of the construct of interest. The correlation between the scores observed on a particular test and true scores (noted $r_{1t} = r_{1(1\cdots k)}$, $k \to \infty$) provides an index of the reliability of this measure. Like any correlation, it can be squared to determine the proportion of observed scores variance explained by the true scores.

$$r_{1t}^2 = \frac{\sigma_T^2}{\sigma_{X_1}^2}$$

where σ_T^2 is the true score variance and $\sigma_{X_1}^2$ is the observed scores variance.

Since the measurement error, *E*, is assumed to be random, it does not correlate with anything else and it's also possible to write

$$\sigma_{X_1}^2 = \sigma_T^2 + \sigma_E^2$$

On the face of it, these relationships might not seem very useful as true scores, errors and their respective variance are unknown and researchers only have access to observed scores. With a few extra assumptions, in particular that the average correlation between a given measures and all potential measures is equal to the grand average of all correlations between potential measures ($\bar{r}_{1j} = \bar{r}_{ij}$), it is possible to

1 Classical test theory is a loosely defined set of models and approaches sharing some important results. "Classical" models are contrasted with "modern" approaches, especially those based on item-response theory. While they do have some advantages, those measurement models will not be considered here because they are much less common in design-related fields and typically require much larger participant samples to be useful. In any case, some of the issues raised in the second part of the chapter would also need to be addressed for these models.

show[2] that

$$r_{1t} = \sqrt{\bar{r}_{ij}}$$

The unobservable correlation between observed scores and the hypothetical sum of scores on all potential measures can therefore be reformulated as the average of the correlations between all possible pairs of observed measures. This correlation can in turn be estimated by the average correlation between any numbers of actual measures:

$$r_{1t} = \sqrt{r_{11}}$$

This last quantity (r_{11}) is the reliability coefficient. In addition to the interpretations mentioned above (correlation between observed scores and the hypothetical true scores, proportion of true score variance in observed scores), the reliability coefficient is used in many results from classical test theory. For example, it can be used to predict the reliability of a test composed of several measures:

$$r_{kk} = \frac{k\bar{r}_{ij}}{1+(k-1)\bar{r}_{ij}}$$

where k is the number of component measures in the new test[3]. A special form of this equation, for k = 2 is

$$r_{kk} = \frac{2r_{12}}{1+r_{12}}$$

It is known as the split-half measure of reliability. Under the assumptions of the model described above, the same formula can also be used to derive the following expression:

$$r_{kk} = \frac{k}{k-1}\left(1-\frac{\sum\sigma_i^2}{\sigma_y^2}\right)$$

where σ_i^2 is the variance of each individual measure/item and σ_y^2

2 The model presented here is called the "domain-sampling model". It is only one of several true score models that can be used to reach the same conclusions, with slightly different sets of assumptions. It is only presented here to help interpret reliability coefficients and introduce α. See Nunnally (1967) for more details on the derivation, other models and relevant references.

3 All the results presented here hold just as well for single items as for multi-item tests. The individual "measure" used to create the new test can therefore be a single item or a group of items, i.e. a set of shorter tests.

is the variance of the sum of these measures.

This is one of the expressions of coefficient α. This coefficient has proven extremely popular and is the most used measure of reliability in applied research by far (Hogan, Benjamin & Brezinski, 2000), probably because it can be directly computed on any test data, without requiring any arbitrary split or the development of new, alternate tests.

However, in spite of its ubiquity coefficient α is in fact frequently misinterpreted and suffers from a number of drawbacks. First, unlike what is often believed, α is not an index of unidimensionality. In fact, all the results presented above still hold mathematically for factorially complex measures or items. Items used in a test can reflect different constructs or be composites of several factors and still have high reliability, as long as the average correlation with the sum of all items is the same. Of course, such a composite measure is much more difficult to interpret and would be less interesting to researchers but, in the true score model, neither reliability nor α depend on unidimensionality. Sijtsma (2009) presents fictional data with very high alpha for bi- or tri-factor questionnaires and explains how to generate data with several clusters of items and an arbitrarily high α.

Additionally, α does not only depend on the internal consistency of the variables (i.e. the magnitude of correlations between them) but also on the number of measures (e.g. items) used in a composite scale. While this property is often presented as a problem, it does make sense. Summing or averaging several measures does actually produce a more stable and reproducible measure. Lengthening scales is a common technique to increase the reliability of a measurement instrument.

Finally, the assumptions underlying the equation of α to reliability (called "essential tau-equivalence") are often not met in practice. If the measures used are not essentially tau-equivalent (i.e. true score variance is not the same for each item), α is only a lower bound to reliability and not necessarily the best one (Sijtsma, 2009).

Another issue with α lies in the way it is used in the applied literature. Reliability is thought as some sort of test should pass and α is evaluated by comparing it to somewhat arbitrary criteria (Lance, Butts & Michels, 2006). Consequently, the development of a measurement tool (especially multi-item self-report scales) typically involves selecting items to reach some threshold and declare the tool reliable. While based on a laudable concern for the quality of new measurement tools, this approach can have undesired effects.

The first of these effects is the tendency to consider reliability as a fixed property of a measurement tool. A simple look at the definition of reliability given above immediately reveals that it is not the case. Even if the magnitude of random error is assumed to be constant across observations, reliability estimates in a given sample will depend on the range of true scores present in this sample. When measuring

individual traits like intelligence, a random sample of the population of interest can be expected to provide a representative range of abilities and therefore a reasonable estimate of scores' reliability across the population. This estimate is however not applicable when working with a subsample of the original population (Feldt & Brennan, 1989). In educational measurement, one such situation arises in research using admission test results. Students admitted to a particular higher education institution will have higher scores than those who were turned down and exhibit a narrower range of scores than the broader population. The reliability of these scores will therefore be lower than that observed in validation studies across the whole population of potential test takers (e.g. secondary school graduates or young people of the same age).

Furthermore, data collected in various fields suggests that error variance itself also depends on the population considered. Vacha-Haase (1998) developed a specific meta-analytic approach called "reliability generalization" to relate differences in reliability and relevant demographic or methodological variables. For example, Youngstrom & Green (2003) examined 132 studies using the Differential Emotions Scale and found noticeable differences in coefficient α depending on the composition of the sample[4]. Socio-economical status has the largest effect on DES reliability, with higher consistency in ratings from participants with a higher socio-economical status. Reliability estimates from tests' manuals or validation studies therefore cannot be assumed to hold for a study with participants from a different or restricted population.

For this reason, several prominent psychometricians have stressed that reliability is a property of scores and not of tests themselves (Vacha-Haase, Kogan & Thompson, 2000). Heeding to their advice, it seems beneficial to move away from over-optimizing questionnaires to reach a particular reliability threshold, instead making sure to report and interpret reliabilities whenever possible.

This is even more important for design-related research as published reliability estimates very often rely on poorly defined convenience samples of students and cannot be assumed to generalize to any other participant sample. Crucially, even when an effort is made to recruit participants from a broader population (e.g. consumer panels), the range of (true) scores in product-related measures does not only depend on the participants' population but also on the choice of products included in the study. There is no reason to assume that

4 Youngstrom & Green (2003) only considered the *trait* version of DES measuring how frequently research participants experience each emotion. Trait affect is thought as a stable characteristic of the participants, much like personality traits and not as a transient *state* like the emotions measured in this thesis.

differences in perceived usability or user experience of the same magnitude can be observed within different product categories. It is also clear that variance in research studies with products deliberately selected to elicit widely different user experiences will be higher than in a comparison between two relatively similar prototypes in a product test at a late stage in the design process.

The second detrimental effect of the "dogmatic" view of reliability is that it obscures some of its practical consequences. Often, computing α is approached as a "black-box" procedure; something that must be done because textbooks' authors claim that reliability is important and reviewers want to see some coefficient reaching a threshold to be satisfied that a questionnaire "is reliable". The whole exercise is therefore perceived as a purely academic concern of little relevance for practitioners. In fact, measurement error and reliability have profound effects on usual statistical analysis procedures (Liu & Salvendy, 2009).

The impact of measurement error on statistical power is rarely mentioned in introductions to reliability and psychometrics. The issue was somewhat controversially discussed in the 1970s following Overall & Woodward (1975) revelation of an apparent paradox in the relationship between reliability and power. Under some assumptions, increased reliability of individual scores results in reduced power for significance tests involving group means. The source of the controversy lies in the definition of reliability presented earlier: the value of the reliability coefficient depends on two different components, true score variance and error variance (or equivalently total observed variance and either true score or error variance). In fact, there is no functional relationship between reliability and statistical power but there is a direct link between error variance and power, as already established by Sutcliffe (1958) and Cleary & Linn (1969). If changes in reliability do in fact result from changes in measurement error, better reliability mechanically increases statistical power.

Even if the confusion was convincingly resolved by the end of the 1980s (Williams & Zimmerman, 1989; Williams, Zimmerman & Zumbo, 1995), measurement error and statistical power are rarely integrated with reliability traditionally presented in the context of statistical tests and individual differences whereas texts on experimental research methodology implicitly assume perfect reliability of individual scores.

7.2. Fundamental issue in product experience measurement

The confusion around reliability and power of significance tests for means touches upon a major difficulty facing researchers and practitioners dealing with product-related measurement: the definition of the object of measurement and multiple sources of error variance.

As noted before, most of the concepts and statistical tools in psychometrics are traditionally discussed in reference to personality or intelligence assessment and educational measurement. In a typical psychological testing situation, a respondent (or test-taker) has to complete a number of tasks or answer a number of questions and the outcome is a small set of scores or numbers thought to quantify some stable characteristics of the test-taker in question. Measurement error results from inconsistencies between items or test sessions whereas differences between people are desirable as they potentially represent the quantity of interest to the researchers. Indicators like coefficient α and test-retest correlations allow the quantification of this error and their use and interpretation is based on the assumption that each participant provides one data point for each condition (i.e. each item, each testing session, etc.)

Design-oriented measures are fundamentally different because they typically quantify product attributes, and not person attributes. What researchers and designers alike are interested in is the impact the product has on its users and not simply stable characteristics of the users. Comparisons between products therefore involve at least two sources of variance beyond the product itself: measurement error in the individual scores and sampling error associated with differences between participants.

This conceptual difficulty manifests itself on a very practical level when computing a reliability coefficient. Published research reports on product-related measurement including reliability data are often elusive on the way the data was processed but obtaining a single meaningful reliability estimate is not trivial in the context of typical experimental designs for product tests. For example, a common approach is to have a number of participants use each product in turn and report their experience about each product (within-subject design).

Armed with such a data set and any common statistical package, there are several ways one could obtain a reliability estimate (say α). A simple one is to treat the whole data set as one big questionnaire, ignoring the fact that each item is in fact repeated several times (one for each product). Even before considering its correctness and interpretation, this reliability estimate suffers from a major drawback: the total number of ratings per participant is several times the actual

number of items. As explained before, coefficient α – or indeed any estimate of total score reliability for multi-items scales – is correlated with the length of the scale and would therefore overestimate the reliability of each individual product rating.

Another approach is to consider ratings for each product separately and compute several reliability estimates. These estimates are actually quite reasonable. One drawback is that this approach does not produce a single reliability estimate but as many as there are products in the study[5]. Yet another approach would be to average ratings for each participant across products, thus falling back to a data set with a single column per item and a single row per participant. Interestingly, the data could just as well be averaged in columns, yielding a single set of ratings per product. As far as we can tell, none of these approaches seem to be used in the literature.

Finally, a tempting approach is to simply "pool" or concatenate all ratings ignoring the structure of the data set. In this setup, each row contains a single rating for each item (i.e. the rating for a specific participant x product combination). Superficially, the data set resembles the results from traditional psychometric studies, with one item per column and one observation per row. Even if the ratings in different observations are not independent anymore, this approach appears to be quite common. Unfortunately, values of coefficient α computed on such a data set are seriously overestimated and do not typically reveal anything interesting to potential users of product-related measures.

These issues can be illustrated with simple numerical examples. All of the mock data sets discussed below correspond to a study in which a four-item questionnaire is administered to 3 participants, each rating 3 products. In the first example (presented in tables 7.1.1 and 7.1.2), the questionnaire only measures some fixed characteristic of the participants. All three products (A, B and C) have the same mean rating on the scale. If the items ratings are simply concatenated (ignoring the lack of independence between observations), α is .98.

Table 7.1.1. *Example 1: Item data for a questionnaire with no product effect.*

	Product A				Product B				Product C			
Items	A	B	C	D	A	B	C	D	A	B	C	D
Person A	1	2	1	2	1	2	1	2	2	1	1	2
Person B	2	3	2	3	2	3	2	3	2	3	2	3
Person C	3	4	3	4	3	4	3	4	3	4	3	4

5 They could however presumably be averaged to obtain a single figure.

This example shows that when treating the data set in this way, α can be very high even if there is no common product-related covariance at all between the items in the questionnaire. In this case, α depends mostly on the number of items and on the ratio between participant-related variance and item-related variance. Alpha, or indeed any internal consistency estimate, has no relationship with the reliability of the scores understood as measures of some attribute of the products tested.

Table 7.1.2. *Example 1: Descriptive statistics for scores with no product effect.*

Statistic	Mean
Per-product scores	
Product A	2.5
Product B	2.5
Product C	2.5
Per-participant scores	
Participant A	1.5
Participant B	2.5
Participant C	3.5

This might seem somewhat obvious as the association between the scores and the products is purely arbitrary and these data are in fact similar to the type of ratings that could be obtained if a personality test with a high short term test-retest stability was administered repeatedly, randomly labeling each repetition "product A", "product B" or "product C". It is however important to understand that design researchers reporting and commenting reliability coefficients or correlations in the absence of differences between products might be dealing with just this type of data. That is, high apparent internal consistency or correlations (between items, questionnaires or with measures of physiological activity or behavior) do not prove that the scores reveal anything at all about the products tested when they are computed on concatenated data.

When concatenating data from several observations, it is perfectly possible to observe high reliability coefficients even if the only systematic source of variance is at the person's level. One plausible scenario generating this kind of data could be that the participants differ in their understanding of the questionnaire or that they are broadly positive or negative towards all products depending on their mood on the day of the test. While in such a study the ratings are ostensibly about the product or condition, they only measure personality traits or current state of the participants.

Table 7.2.1. Example 2: Item data for a questionnaire with weak product-related variance.

	Product A				Product B				Product C			
Items	A	B	C	D	A	B	C	D	A	B	C	D
Person A	1	1	1	1	1	2	1	2	2	2	2	2
Person B	2	2	2	2	2	3	2	3	3	3	3	3
Person C	3	3	3	3	3	4	3	4	4	4	4	4

While this example makes an important point, it represents an extreme case. Such measures are probably rarely encountered in practice, at least with self-report user experience questionnaires. After all, empirical papers on such questionnaires typically include at least some differences between different products. A more interesting scenario is presented in table 7.2.1 to 7.3.2. In this fictional study, two questionnaires with the same format are used by three participants to rate three different products. For both questionnaires, scores vary systematically depending on participants *and* on products.

Table 7.2.2. *Example 2: Descriptive statistics for scores with weak product-related variance.*

Statistic	Mean
Per-product scores	
Product A	2
Product B	2.5
Product C	3
Per-participant scores	
Participant A	1.5
Participant B	2.5
Participant C	3.5

This situation is pretty typical for all types of user experience measures. As expected, different products elicit different experiences but the scores also differ from participant to participant. This participant effect might reflect differences in personality, mood when testing the products, understanding of the questionnaire or response sets. For example some participants might not be comfortable expressing strong emotions in relation to products and generally use lower ratings, others might have a broadly positive outlook on the product category and provide generally positive ratings across products, etc.

In the examples presented here, the participant and product effects are additive. This means that participants use a different "baseline" but react similarly to each product and there is no interaction between participants and products. The key difference between the two

questionnaires lies in the respective size of the product and participant effects.

In the first questionnaire (presented in table 7.2.1 and 7.2.2), differences between products are modest and the mean score difference between the most extreme products is only 1 point (expressed in the same unit as the original rating format). The differences in mean scores between participants are bigger, with 2 points between the participant reporting the lowest level of experience and the one reporting the highest.

Table 7.3.1. *Example 3: Item data for a questionnaire with strong product-related variance.*

	Product A				Product B				Product C			
Items	A	B	C	D	A	B	C	D	A	B	C	D
Person A	1	1	1	1	2	2	2	2	3	3	3	3
Person B	1	2	1	2	2	3	3	2	3	4	3	4
Person C	2	2	2	2	3	3	3	3	4	4	4	4

In this fictional study one of the questionnaires is more sensitive to participant characteristics whereas the other is strongly influenced by product-to-product differences. In both cases, α is very high (.98) and it does not differ from one questionnaire to the other. In design research and product tests however, participant effects are a source of error and these two questionnaires are far from being equally useful.

Table 7.3.2. Example 3: Descriptive statistics for scores with strong product-related variance.

Statistic	Mean
Per-product scores	
Product A	1.5
Product B	2.5
Product C	3.5
Per-participant scores	
Participant A	2
Participant B	2.5
Participant C	3

One way to understand these examples is to turn back to the definition of reliability and the derivation of α exposed earlier. Under the assumptions of classical test theory, α has been shown to be an estimate of the reliability of a measure, defined as the correlation between the observed scores and the underlying hypothetical true scores. It has also been established that α can be interpreted as the square root of the average inter-item correlation or as the mean

of all split-half correlations. In all these interpretations, α is simply a coefficient of correlation (or a simple function of a correlation coefficient).

Correlation coefficients are a natural measure of the strength of a linear relationship between two variables and are used extensively to assess the association between two variables. Their interpretation is however much more complex than often realized, and the magnitude of a correlation depends on many other factors than the strength of the relationship between the variables. One of these difficulties of interpretation is called "Simpson's paradox".

When aggregating data from several groups, the correlation between two variables over the whole data set can be very different from the correlations within each group.

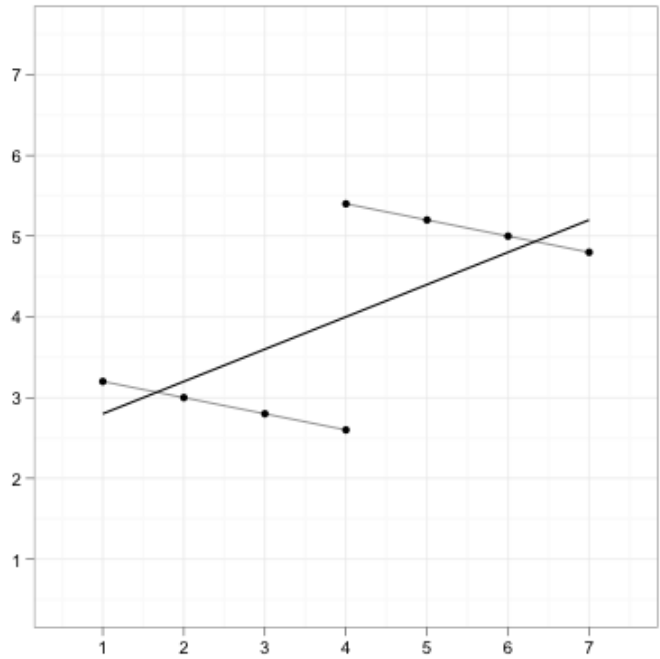


Figure 7.1: *Simpson's paradox with continuous variables. The lines represent linear trends, within the two groups (thin line) and over the whole data set (thicker line). In the aggregated data set, the perfect (negative) linear relationship within each group is obscured by the difference between the two groups and replaced by a relatively strong positive correlation (r = .66).*

In the data-set represented in figure 7.1, the relationship between the two variables is inverted when considered at the group-level, compared to the aggregated data set. There is a perfect negative correlation between scores within each group and a strong positive correlation over the whole data set.

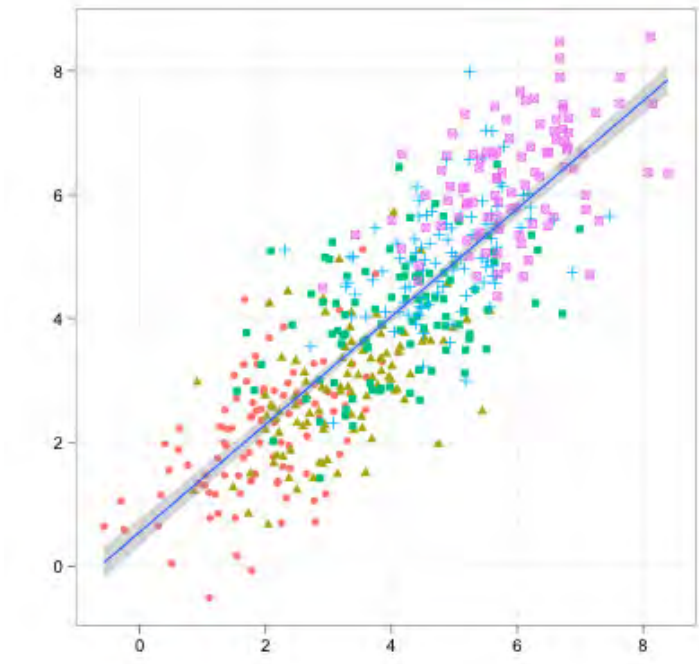


Figure 7.2: *Another illustration of Simpson's paradox. Color/shapes represent different groups. In this example, the correlation over the whole data set is .82, correlations within the groups range between .43 and .55 and the correlation between group means is .99.*

The problems with correlation computed on aggregated data are however by no means restricted to such extreme situations. In the data set represented in figure 7.2, the correlation between the measures is positive at all level of analysis but the magnitude observed on the pooled data represents neither the within-group nor the between-group level of correlation.

7.3. Generalizability theory

Generalizability theory (Brennan, 2001; Shavelson & Webb, 1991) is a framework that can be used both to better understand the issues touched upon in the previous section and to properly address reliability issues in user experience measurement. The central concept of generalizability theory is that each score or measure is a single sample from an infinite universe of acceptable measures.

For example, items in a questionnaire represent a sample of a larger set of acceptable items measuring the same attribute. Typically, researchers are not specifically interested in the score on

the specific items used but would accept many other similar items as long as they measure the same quantity. In generalizability theory, this (hypothetical) set of potential items is called the universe of generalization, and reliability (or generalizability) is conceptualized as the accuracy of the generalization from observed scores to universe scores (the hypothetical average score across all acceptable items).

Formally, the score obtained by a participant *p* on a item *i* is

$$X_{pi} = \mu + \nu_p + \nu_i + \nu_{pi,e}$$

$$\begin{aligned} X_{pi} = & \mu + \\ & \mu_p - \mu + \\ & \mu_i - \mu + \\ & X_{pi} - \mu_p - \mu_i + \mu \end{aligned}$$

μ is the grand mean across all participants and items, μ_p is the participant's difference score and μ_i is an item's offset. $\nu_{pi,e} = X_{pi} - \mu_p - \mu_i + \mu$ is a residual factor, capturing all other sources of variance. Except the grand mean, all effects have a distribution with means 0 and a specific variance. For example $E_p(\nu_p) = E_p(\mu_p - \mu) = 0$ is the mean of the participant effect and $\sigma_p^2 = E_p(\mu_p - \mu)^2$, its variance, represents the magnitude of the differences between participants.

Even if the formalism is a bit different than the classical test theory presented at the beginning of this chapter, the underlying idea is very close to the domain-sampling model. Each effect is associated with a variance component. The variance component for the item effect represents the error in generalizing from a single item to all potential conditions in the universe of generalization.

The force of generalizability theory is that it becomes possible to introduce several sources of error and consider them concurrently. Whereas in a classical setting, test-retest reliability and internal consistency would be assessed separately, they can be combined in generalizability theory. The corresponding score decomposition is

$$X_{pio} = \mu + \nu_p + \nu_i + \nu_o + \nu_{pi} + \nu_{po} + \nu_{io} + \nu_{pio,e}$$

In generalizability theory, sources of error variance are called "facets". This model includes two facets (items and occasions), a participant effect and the interactions between them. Including different facets allow researchers to define the universe of scores they intend to generalize to. In practice the corresponding variance components are

$$
\begin{aligned}
X_{pio} = {} & \mu + \\
& \mu_p - \mu + \\
& \mu_i - \mu + \\
& \mu_o - \mu + \\
& \mu_{pi} - \mu_p - \mu_i + \mu + \\
& \mu_{po} - \mu_p - \mu_o + \mu + \\
& \mu_{io} - \mu_i - \mu_o + \mu + \\
& X_{pio} - \mu_{pi} - \mu_{po} - \mu_{io} + \mu_p + \mu_i + \mu_o - \mu
\end{aligned}
$$

estimated using an ANOVA framework and can be used to understand the effect each facet has on measurement reliability.

While the emphasis is on the estimation and interpretation of variance components and measurement error, it is also possible to define several coefficients to index the precision of a measure. In the two-facet *item x occasion* design, the generalizability coefficient (analogous to the reliability coefficient in classical test theory) is defined as

Importantly, generalizability theory makes a distinction between G-studies and D-studies. G-studies are used to estimate the magnitude of the different sources of error involved in a particular measurement process whereas D-studies use these estimates to predict the effect of changes in the measurement process. Concretely, in the formula for

$$
E\rho^2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{pi}^2 / n_i + \sigma_{po}^2 / n_o + \sigma_{pio,e}^2 / n_i n_o}
$$

the generalizability coefficient, n_i and n_o need not be equal to the number of items and occasions in the G-study but can be modified, for example to predict how extra items might impact measurement error.

All the examples introduced until now pertain to the measurement of personal attributes but generalizability theory is by no means limited to this type of situations. In fact, one of its key advantages for user experience research is that it offers an explicit framework to define true (or universe) score and measurement error. In the score decompositions, the participant effect ($\nu_p = \mu_p - \mu$) is expressed in the same way as the item ($\nu_i = \mu_i - \mu$) or occasion effects but the corresponding variance components do not all contribute to error variance in the formula for the generalizability coefficient.

Conceptually, this formalization offers a key to the understanding of the difference between the various hypothetical measures presented earlier. In fact, a study in which several participants rate several products

with a multi-item questionnaire can be understood in generalizability theory as a two-facet crossed design. However, in design research it is often the product or design and not the participant that should be considered the object of measurement. A measure with no product-related variance (example B1-B2) should properly be considered as having zero reliability. In a G-study based on these data, σ^2_{prod} would also be null and the corresponding generalizability coefficient would consequently also be equal to zero. Participant-related variance (which might very well account for high correlations between items as demonstrated before) is properly considered error variance and does not appear on the denominator of the generalizability coefficient.

To illustrate the types of conclusion enabled by generalizability theory, let us consider a study mentioned before: The comparison between PrEmo and Geneva Emotion Wheel (GEW) ratings after using a coffee maker and an alarm clock (chapter 3). The two instruments (PrEmo and GEW) will be analyzed separately. All emotion descriptors (words or animations) of the same valence are grouped to form a positive and a negative emotion scale for each instrument. Since all participants (noted *p*) used both products (noted *d*) and rated them with the same set of emotions (noted *e*), the study design is $d \times p \times e$. Table 7.4 presents estimates of the various variance components involved based on the data collected in the study[6].

6 All generalizability theory analysis were performed using Brennan's GENOVA (see Brennan, 2001).

Table 7.4. *Variance components for various questionnaires used in the coffee machine/ alarm clock study (G-study).*

	Negative emotions		Positive emotions	
	GEW	PrEmo	GEW	PrEmo
$\hat{\sigma}^2_d$	0.1689	0.0889	0.1165	0.0729
$\hat{\sigma}^2_p$	0.0592	0.0122	0.3140	0.0328
$\hat{\sigma}^2_e$	0.0509	0	0.3083	0.0404
$\hat{\sigma}^2_{dp}$	0.1100	0.0661	0.1668	0.0646
$\hat{\sigma}^2_{pe}$	0.0563	0.0199	0.1670	0.0658
$\hat{\sigma}^2_{de}$	0.2781	0.0335	0.0487	0.0202
$\hat{\sigma}^2_{dpee}$	0.8480	0.3490	1.2842	0.2923

The absolute values of the variance components can't be directly interpreted but the proportion of total variance for each component indicate how important each source of error is. For both questionnaires and both scales, the biggest component is $\hat{\sigma}^2_{dpee}$, between 50 % (PrEmo/positive emotions) and 61% (PrEmo/negative emotions) of the total variance for the relevant scale. It represents error variance that isn't specifically related to any of the facets included in the design together with the three-way interaction between product, participant and emotion (in any G study, the highest level interaction and error associated with facets not included in the design are confused in the residuals). For PrEmo scales, the product effect is the next biggest component. For both GEW scales, other components (participant and emotion effects for the GEW positive emotions scale and product x emotion interaction for the GEW negative emotions scale) are also bigger than the product effect.

The results from the G-study can also be used in a D-study to see how changes in the number of participants and emotions influence the reliability of the product scores (table 7.5). Several conclusions can be drawn based on these results. First, adding participants or items improves the reliability of the scale. Quite obviously, a single rating

can't be readily generalized to other participants or items. The average (or sum) score over several conditions is more generalizable because, as with any other mean, it is more stable and closer to the population value (or universe score).

Table 7.5. Generalizability coefficients for mean scale scores per product based on various scenarios for the number of items/emotions and participants (D-study).

		Negative emotions		Positive emotions	
Emotions	**Participants**	**GEW**	**PrEmo**	**GEW**	**PrEmo**
1	5	.26	.43	.26	.44
5	5	.60	.72	.55	.72
10	5	.72	.79	.65	.78
1	10	.31	.54	.38	.57
5	10	.67	.81	.69	.82
10	10	.78	.87	.77	.86
1	20	.34	.62	.49	.66
5	20	.71	.87	.79	.88
10	20	.82	.92	.86	.92

Second, there is a trade-off between the different facets: reliability or generalizability can be improved in different ways. Thus, a scale with more emotions (items) needs fewer participants to achieve a given level of precision. In this study, a product test with 10 participants and single-item scales would have large measurement error and dismal levels of generalizability. Using PrEmo five-emotion scales however

makes the precision of the measures obtained with this relatively small sample acceptable.

Third, the increase in generalizability when adding extra conditions levels off. For example for the PrEmo positive emotions scale, 5 emotions represent a dramatic improvement over a single item but the (predicted) average over 10 emotions is only slightly more generalizable than the score based on a five-emotion scale.

Together, these findings illustrate the practical impact of measurement reliability on the cost and time needed for product tests. Generalizability can be used to optimize these tests by pointing to the best ways to gain precision at a reasonable cost.

7.4. Conclusion

This chapter highlighted the link between measurement reliability and statistical power before describing some difficulties in applying these notions to within-subject experiments and briefly describing a framework that would be useful to assess and compare the reliability of user experience measures.

While many of the issues raised here apply to many kinds of research, they are especially important for design-related research, as many of the "tricks" available to compensate poor measurement reliability in experimental research (e.g. using more extreme stimuli or many trials in each conditions) are not always practicable when working with actual products. Improvement of measurement reliability can also ensure more efficient assessment of the user experience of various products by reducing the number of participants required to obtain a given level of precision, an issue that is particularly relevant to practitioners working under stricter time-constraints, often without access to a cheap pool of participants like students.

8. On Validity

A common definition of measurement validity is the extent to which a measure reflects what it purports to measure. Validity is therefore a key aspect in the development of new measurement processes and the choice of measures, both in academic research and user experience tests. This chapter will provide an overview of the major conceptions of measurement validity, drawing primarily on the psychometrics literature and discuss their applicability to applied research on design and user experience. Empirical results (especially from chapters 3 and 6) will then be revisited to examine how they speak to the validity of the different measurement techniques used in the thesis. Finally, the notion of measurement validity will be used to shed some lights on the differences and similarities of the various families of emotion measures reviewed in chapter 2 and identify some important issues in the way those are usually discussed in the applied literature.

8.1. Kinds of validity

Several distinct conceptions of validity have been advanced in the psychometrics literature. A common way to summarize this literature distinguishes, in chronological order, between criterion validity, content validity, and construct validity, each of these view of validity appearing after serious problems in the previous one become evident and culminating in a unified view of validity combining many aspects or kind of validities (e.g. Messick, 1995). The overview presented here draws extensively on Messick, Kane (2001) and Zumbo (2007). The – rather different – perspective developed by Borsboom, Mellenbergh & van Heerden (2004) will also be presented and inform the discussion of the various types of emotion measurement tools.

8.1.1. Criterion validity

The first kind of validity discussed here, criterion validity, is probably the most intuitive approach to validity. From this perspective, the validation of a new or proposed measure is based on the equivalence between this measure and some other established measure of interest, the criterion. Validation simply becomes a comparison between a new

measure and a reference[1]. Of course, this approach presupposes the availability of a reasonable criterion and often achieves little more than moving the validity problem one step further to the measurement of the criterion itself. Often, psychological variables are not easily measurable and new measurements are devised precisely because none of the available measures is fully satisfactory.

Still, in many applied fields, the idea of an association between test scores and later outcomes makes a lot of sense, and indeed applications of tests in educational or industrial settings were instrumental in the development of criterion validity. For example, a common goal for admission tests for higher education institutions is to select the students that would be most likely to succeed and to rank candidates according to their ability to complete their studies.

Similarly, criterion validity would seem very relevant to measures collected during product tests. If designers and researchers are interested in perceived usability/satisfaction or product-related emotion in the first place, it is because these responses are widely thought to influence our willingness to buy and use specific products. The user experience measured in a short, lab-based product test is seldom a goal in and of itself. At the very least, measures of user experience obtained in a product test are intended as a proxy for an evaluation of the experience users would have after using the product for themselves and not only within the restricted context of a user research effort. Looking at the correlation between pre-launch assessment of product-related emotions and success on the market, sustained use or experience outside of the lab would therefore seem to be an excellent way to establish the validity of a measure of emotion for product design.

Unfortunately, this type of associations is very difficult to assess in practice because there is a considerable time between the measurement and the final outcome and many other factors can have an impact on this outcome. Additionally, since criterion validity is typically assessed with correlation coefficients, validity will depend on the specific population used to evaluate it and many well-known effects can distort apparent correlations. For example, student selection during admission (with the new measure or by some other means) is likely to strongly attenuate any empirical association between the test and a criterion. Since only a limited and rather homogeneous group of students is admitted, any criterion measured on this group of students will have a severely restricted range and therefore a reduced correlation

1 The name "criterion validity" is closely associated with educational and psychological measurement but broadly similar ideas also appeared in neighboring fields such as affective computing or human factors under a different terminology (e.g. discussion of "ground truth" or "gold standard").

with any other variable. Similarly, from the many conceivable designs or actual prototypes developed, only a fraction will ever become finished products. Whether some type of formal user experience testing is a part of the design process or not, they will certainly not be selected randomly and should even ideally be the best possible designs according to the criterion of interest, thus reducing the empirical relationship between that criterion and any other variable, when estimated on those designs that were made into an actual product.

A more fundamental problem lies in the use of correlations to define criterion validity. In social science, all variables can be expected to be empirically correlated, if only moderately, leading to the unsatisfactory conclusion that any variable has some limited degree of validity as a measurement of just about anything (Borsboom, et al., 2004). What the criterion should be is not entirely clear either. In design-oriented research, economic criteria are of course relevant for many practitioners but even those are not trivial to define and measure (number of units sold? profits?) It is also obvious that user experience or even design in general is far from being the only factor influencing a product's success. Competition, marketing, and price are but a few of the other variables that can obscure the purported link between a great experience and commercial success. Conversely, a design can be deemed to be successful because it possesses a certain aesthetic appeal, satisfies a particular need or delivers a great user experience even if it fails to sell well. That a given product is not economically successful cannot automatically be taken as evidence that it is ugly or has a poor usability or user experience. In fact, using these variables as criteria substitutes predictive usefulness for measurement validity. Even if user experience does indeed contribute to a product success and measures of experience do predict it, this success is obviously not the same thing as the experience itself.

8.1.2. Content validity

The notion of content validity is an attempt to address some of these problems by replacing correlation between a measure and a criterion with expert judgment about the representativeness of a measurement instrument. It is easiest to understand in reference to tests assessing domain-knowledge, for example in education or recruitment. Thus, a test to select employees to fill a particular position should be representative of that position, i.e. reflect all knowledge and skills needed to successfully carry out the duties associated with it. A test that only assesses a small part of these skills can fail to rank highly the most promising prospective employees and provide a poor basis of decision.

Content validity is a little more difficult to extend to the measurement of emotions. One aspect of measurement procedures (especially multi-scales self-report questionnaires) that could fall under this label is the representativeness of the set of emotions or affective states included. Thus, Barrett & Russell (1999) or Larsen, Norris, McGraw, Hawkley, and Cacioppo (2009) stress than measuring a single dimension (i.e. valence or hedonic tone) can produce a distorted view of a person's affective state. The argument is that measurement tools should cover the whole (two-dimensional) space of affect, whether it is with multi-item scales (e.g. adjective ratings), with several single-item scales (e.g. self-assessment manikin) or with single-item instruments like the affect grid or the evaluative space grid. Based on data reduction analyses performed on ratings of the semantics of frequent emotion terms, Fontaine, Scherer, Roesch, and Ellsworth (2007) go one step further and argue that evaluation/pleasantness and activation/arousal are not enough to fully describe affective states and that two other dimensions (potency/control and unpredictability) should also be included. Validation of the content of emotion measurement instruments taking these findings into account would therefore presumably involve checking whether the instrument includes items reflecting all combinations between these four dimensions.

The key idea behind content validity, sampling the domain being assessed, can also be applied to the processes involved according to relevant theories in that domain. The multi-componential view of emotion evoked in chapter 2 would therefore lead to an instrument involving all the main components of emotion: subjective experience, bodily activation, facial expression, behavior, and appraisal. A major difficulty in the application of the notions of sampling and content validity to emotion measurement is that they rely crucially on a clear definition of the boundary of the domain to be assessed. Given the broad definitions and wide differences in the field, the "emotion" concept does not seem very useful in that respect. In fact the distinction and boundary between "affective" and "non-affective" is far from obvious and slightly controversial, both in terms of states or contents (are surprise or boredom emotions?) and in terms of processes or systems (are facial expression an integral part of affect or only loosely associated with it? Are feelings necessary for emotions?) and researchers diverge in their choices on the issue.

8.1.3. Construct validity

Construct validity is another attempt at addressing the difficulties inherent in criterion validity. Since a good criterion often remains elusive and validation would anyway be a moot point if one were

available, construct validation aims at "bootstrapping" psychological measures by replacing correlations between a measure and some external reference with the study of correlations *between* different measures. The multitrait-multimethod matrix (Campbell & Fiske, 1959) is a simple technique based on this notion. Such a matrix (abbreviated MTMM) results from the measurement of several attributes or "traits" with several instruments or "methods". Ideally, the various methods used to measure one trait should be independent and as different as possible but have high correlations between themselves (convergent validity) while the correlations between different traits, whether measured with the same method or other methods should be as low as possible (divergent validity). In a MTMM, the relationship between various measures, and not the association with a reference or criterion, is therefore interpreted as evidence for their validity. Nomological networks are a somewhat more elaborate way to investigate construct validity. A nomological network specifies the relationships between different measures in the form of expected correlations (positive or negative) between them. Validation involves checking the empirical fit between the predicted network and observed correlations between variables.

For emotion research, one obvious application of the notion of construct validity is a comparison between measurement tools based on the various components presented in chapter 2. If, as expected from the most influential definitions, emotion results from the coordination of various components or subsystems, measures tapping these different components should exhibit strong correlations over a range of emotion-eliciting conditions. Unlike correlations between slightly different self-report scales, such a finding would be non-trivial and encouraging regarding the validity of the measures used. Empirically, however, observed correlations between measures of the various components of emotion tend to be quite low (Bonanno & Keltner, 2004; Mauss, McCarter, Levenson, Wilhelm & Gross, 2005). Other authors suggest that some emotions are unconscious, completely decoupling subjective experience from other components, including visceral reactions and approach/avoidance behavior (Berridge & Winkielman, 2003). Obviously many technical and methodological difficulties can account for these disappointing results but they still have important consequences for the measurement of emotion in research and practice. Some researchers have also offered other interpretations, suggesting for example that the subsystems involved are only loosely coupled or that response coherence might only be a characteristic of folk concepts of emotion, therefore not necessarily present in non-prototypical emotional episodes (Russell, 2003). Describing emotions as multi-componential responses would therefore not imply any commitment to a particular level of correlation between the different components.

Studies directly addressing this issue in the context of design research or human-computer interaction are scarce and research reports including measurement of several components of affect (e.g. self-report and physiology) do not always allow any clear conclusion about the magnitude of the correlations but those that are reported also tend to be quite modest (e.g. Mahlke & Thüring, 2007). Here again, a number of technical issues (reliability of the various measures, especially physiological ones, choice of products tested and dichotomization of some variables, etc.) certainly do attenuate the observed correlation but it seems difficult to argue that the different variables measure a single, coherent process and could be used interchangeably to reach conclusions about user experience.

Another influential conceptualization of validity was developed by Messick (e.g. 1995). While he retains the notion of construct validity, he offers a unified view of validity in which the different types of evidence described until now become "aspects" of a more general validity. He also adds an emphasis on the social consequences of erroneous measurement. Validation efforts should therefore attend to potential detrimental (but also positive) consequences of test use and interpretation. Once again, these ideas are discussed in the context of educational testing but they certainly seem relevant to applied research in design-related fields, considering for example the role of tests and evaluations in design practices and product development and the risk of incorrect decisions due to bias in the measurement process.

Borsboom et al. (2004) offer a starkly different perspective on the validity concept in psychological measurement. Based, in part, on the issues raised earlier when discussing criterion validity, all validation methods centered on correlations (including criterion and construct validity) are deemed inadequate. More fundamentally, current thinking is accused of confusing validation (the different epistemological means to collect evidence of validity) and validity itself (an ontological question). Nomological networks are criticized as "relics" of logical positivism and a failed attempt at thinking about validity without discussing what reality measures refer to. Instead, the focus should be on the causal link between the attribute of interest and the measure, i.e. talking about a valid measure of a given attribute implies that this attribute exists and causes variations in the measure. Validation therefore becomes the specification of the causal mechanism at play in test responses. This deceptively simple idea, it is argued, is much closer to the intuitive understanding of validity held by most researchers, including, incidentally, the definition put forth in the opening of this chapter. Most of the aspects listed by Messick, including the consequences of test use and interpretation, are deemed not to be part of validity at all by Borsboom et al. who instead suggest that they should better be considered part of a looser notion of "overall test quality".

8.2. Empirical evidence

Equipped with the various notions of validity described in the first part of the chapter, it is now possible to review the data presented in the rest of the thesis with an eye toward validation of the measure developed.

In the coffee machine/alarm clock rating study (chapter 3, section 3.1), the correlation matrices between the two questionnaires used can be regarded as multitrait-multimethod matrices[2]: several (group of) emotions were assessed with two different self-report tools, an adjective rating questionnaire (the Geneva Emotion Wheel, GEW) and a non-verbal questionnaire (PrEmo). The highest correlations are those between groups of items measuring the same emotions with different questionnaires (i.e. monotrait-heteromethod correlations in the validity diagonals). The only exception is the correlation between negative emotions in PrEmo and low control/low pleasantness emotions in the GEW, revealing the fact that these emotions (e.g. sadness or guilt) are not covered by the version of PrEmo used in this study. These monotrait-heteromethod correlations provide encouraging evidence of converging validity between the two questionnaires. Most of the heterotrait correlations are not very large, which can be interpreted as a form of discriminant validation (Campbell & Fiske, 1959). Some monomethod-heterotrait correlations are quite significant but negative, which is more likely to reflect the bipolar nature of affective valence than common method variance. Importantly, the pattern of trait interrelationships is the same in all heterotrait "triangles" both in monomethod and in heteromethod blocks, which can also be interpreted as a sign of construct validity (Campbell & Fiske). Overall, the difference in form (adjective ratings vs. non-verbal self-report) between these two measurement tools makes the convergence more significant from a validity perspective, even if both instruments used in this study are self-report questionnaires.

2 Incidentally, the target attribute is not a trait at all but a state induced by the interaction. The validity of any measure of this attribute should therefore also be assessed at the intra-individual level, considering variations in a person's state caused by the use of the product, especially if the measurement instrument is to be used to compare average responses to different products (and not individual differences in response to the same product). The correlation matrices discussed here however reflect variations across persons using the same product and provide a partial view of construct validity, at best. See also the discussion of Simpson's paradox and sources of variance in chapter 7. Despite all this, these results will be discussed using the standard terminology and the word "trait".

The personal navigation device study (chapter 3, section 3.2) used only one emotion measure (a paper-and-pencil variant of PrEmo) and the results cannot be used to build a multitrait-multimethod matrix. The pattern of correlations between the different measures used can however tentatively be interpreted in terms of construct validity, even if no attempt was made to specify a nomological network beforehand. In particular the correlations between emotion ratings and other variables (handiness and originality, perceived usability, pragmatic and hedonic qualities) can be interpreted either on a methodological or on a substantial level. On a methodological level, strong correlations with all other measures could be a sign of a lack of specificity or weak discriminant validity of the emotion scale. On a theoretical level however, these correlations do make sense. Current theories of emotion stress their role in evaluating one's current situation and integrating various sources of information to motivate adaptive behavior and react to opportunities and challenges in the environment. The alternative interpretation is therefore that while usability, originality or aesthetics would be expected to be distinct qualities, they could all be related to emotion understood as a broad evaluation mechanism taking all these qualities into account. The different patterns of correlations in the two parts of the project (see Desmet & Schifferstein, 2010 and chapter 3, section 3.2), including lower correlation between "handiness" and emotion when no actual use is involved and strong correlation with usability when a goal-directed task is carried out, further support this interpretation. However, the limited sample size, the constraints on the experimental design of the study, the lack of formal tests and the post-hoc nature of the interpretation severely limit the reach of these conclusions.

Since both studies were randomized experiments, they also provide evidence of a causal link between the product assigned to each participant and the response recorded by the various measurements. Admittedly, the scope of this evidence is very limited, as it does not provide any insight into the specific causal mechanism involved. For example, the mere fact of a difference between conditions does not establish in and of itself that this difference is the result of affective processes as opposed, say, to some unrelated cognitive process. Still, this evidence is valuable and in fact many publications reporting and interpreting correlations between affective measures in applied research (e.g. physiological signals or facial behavior) do not even provide this minimal level of evidence of product-related variation.

Similarly the difference between products in the self-confrontation studies in chapter 5 point to a causal link between the experimental manipulation and the data collected. This is not very surprising since the instructions and the whole design of the studies make it very clear to the participant that the focus is on the interaction with the products. The result is not entirely trivial however, especially for the

personal navigation device ratings since the between-subject design should prevent explicit comparisons between the products used in the study. Convergence between different participants and differences between devices and over time therefore provide evidence that the data collected during the self-confrontation procedure are in fact causally linked to the interaction with the product even if, once again, their validity as a measure of *affective experience* rests entirely on the instructions themselves. The relationship between the moment-to-moment ratings and the final questionnaires does however provide some additional correlational evidence for the validity of the self-confrontation ratings. Specifically, the link between the peak in the moment-to-moment rating and the final affective ratings matches theoretical expectations and previous results from pain research, and can therefore be interpreted as evidence of construct validity for these ratings.

8.3. Other issues

Despite the large differences regarding the definition of validity and its philosophical underpinnings, all contemporary validity theorists (Messick, 1995; Borsboom, Mellenbergh & Van Heerden. 2004; Zumbo, 2007) do however converge on a number of very generic ideas, namely that substantive theory should inform measurement (albeit not always emphasizing the same type of evidence) and that the same framework should be used to examine the validity of different types of measurement (from ability tests and attitude questionnaires to psychophysiological measurement). These simple yet far-ranging ideas reveal how the findings about the architecture of emotion reviewed in chapter 2 can inform measurement and constitute a strong basis to clarify some thorny issues running through the (applied) literature on the measurement of emotion.

In particular, the notion of a causal link between variation in the attribute and variation in the measure provides a way to think, qualitatively or quantitatively about the validation of different measures (self-report, behavior observation, physiological measurement) in a common framework. Importantly, the traditional distinction between "objective" and "subjective" measures of emotion is not operative in this context; in both cases, the researcher wishes to trace back relatively unproblematic observed data (actual ratings on a questionnaire, changes in electrical properties of the skin) to the psychological or neurological processes producing them. Commonly invoked threats to validity (social desirability, demand characteristics, deception…) can be thought of as alternative causes for the observed changes and empirical research should determine how they impact the different measurement procedures available.

Interestingly, the causes of variations in "objective" measures, especially psychophysiological signals, are not much better defined than the processes underlying self-report. Ironically, the idea to measure affective processes through physiological changes owes just as much to the common sense experience that emotions are accompanied by bodily arousal than to any theory of the mechanisms behind these changes, be it on a functional or on a neurological level. In fact, the most influential conceptualization of the role of the body in emotions, the James-Lange theory, long predated any actual psychophysiological measurement and was entirely based on introspective evidence. Later psychophysiological research mostly adopted a "black-box" empirical approach, relating peripheral changes and functional variables, including experimental manipulations. Often, the meaning of these functional variables ultimately rests on the researcher's intuition or on a pre-selection based on self-report data, and the choice of physiological signals measured depend on convenience and availability. Only recently has research on the neurological systems involved appeared.

Even the suggestion that physiological measures are "objective" and not sensitive to influences like demand characteristics is based on introspection and common sense experience. The distinction between "objective" and "subjective" measures has a strong intuitive appeal and is very easy to grasp. Self-reporting participants must be asked to reflect on the content of their conscious experience and voluntarily report it, whereas facial expressions are constantly "given off" sometimes without us even noticing that we are emitting them. In some settings, they can even be recorded covertly without informing the participants that they are being observed or that the researcher is interested in emotions before the end of the experiment. Similarly, changes in heart rate or skin conductance is not something we feel we can change at will, even if we will see that they are just as sensitive to a range of complex top-down processes and can very well be consciously altered. Conversely, we strongly experience actions like pressing a button or writing down a number as willful, even if it can be shown that unconscious and automatic processes do influence or modulate them as well. The important thing here is that arguments about the validity of psychophysiological measures (at least in the emotion measurement literature) are not based on a clear model of the causal mechanisms underlying variations in this measure or on evidence of the (lack of) influence of any specific threats to validity on this variation; it is based on our intuitive, subjective experience of these influences.

In fact, while several processes are often mentioned throughout the applied literature on the measurement of emotion in human-computer interaction, design or consumer psychology as threats to the validity of self-report, they are never described in detail and the reasons while

they would not impact psychophysiological measures or observation of facial behavior are not specified. Demand characteristics are such a threat. While they are often invoked, sometimes with a reference to Orne's (1962) original paper on the topic (e.g. Levenson, 2003), their consequences for the measurement of emotion are seldom discussed, much less subjected to empirical investigation.

Orne's definition of demand characteristics stems from the fact that subjects in psychological experiments are active *participants* in the study. Experimenters however tend to focus on the experimental manipulation, what *is done* to the subjects, and neglect their active participation in the experiment, what they *do* in the situation. Unlike the results of physics or biology experiments which can be adequately understood by referring solely to the independent variables, the behavior of participants in a psychological experiment is determined by the whole experimental situation, which is always eminently social. This behavior then can be understood as the consequence of two sets of variables: "(a) those which are traditionally defined as experimental variables and (b) the perceived demand characteristics of the experimental situation." (Orne, 1962)

Demand characteristics are first and foremost about independent variables, not about a specific type of response or measures. In fact, the original impetus for Orne's work on the topic came from a pilot study that did not involve self-report at all. Trying to devise a task so boring that participants would refuse to continue doing it, he noticed how powerful the experimental situation itself was, even before any other manipulation – the task would have been the control condition in a hypnosis experiment. The participants in this pilot study did not simply report feeling good about the task to please the experimenter; they actually performed tedious calculations before shredding the results for hours on end. This experiment can be compared to the contemporary "obedience to authority" studies by Milgram (1974)[3]. During these studies, participants were led to inflict increasingly powerful electrical shocks to another participant in what was ostensibly a memory experiment. In fact, the other participant was a confederate and no actual shocks were delivered but in a typical variant of the study, about 65% of the participants would proceed all the way to the end of the experiment, after hearing the confederate complain, scream and finally become silent. In Milgram's case, the power of the experimental situation to bring people to do something they would not otherwise do is the actual variable of interest, not an unwanted artifact but it is interesting to note that here as well the dependent variable is actual behavior, not self-report. Milgram reports that his participants were genuinely distressed by what they were doing; the perceived demands of the experimental situation did

3 I am grateful to Anna Fenko for suggesting this parallel.

not merely exert a superficial influence on the participants but moved them deeply. It's difficult to think that Milgram's participants would have remained completely cold, with no bodily arousal and a frozen face while reporting feeling bad about inflicting pain and possibly killing someone. Simply stated, experimental situations can create genuine affective responses, even very strong ones. There is therefore no a priori reason to assume that unwanted characteristics of these situations could not affect any measure of emotion.

In fact, there is no suggestion in Orne's writings that participants are consciously deceptive. The effect he observed would therefore seem to be mostly driven by an unconscious tendency to conform to the demands of the experiment and be "good subjects". If that is the case, all bets are off and the subjective experience that we cannot steer or control our bodily arousal becomes irrelevant. If, on the other hand, researchers worried about demand characteristics are concerned with conscious, willful deception from their participants, the lack of direct, subjectively experienced control over autonomic systems is no guarantee either. A number of tricks, popularized by fictional descriptions of "lie detectors" in films and television shows, are available to disturb psychophysiological measures, most notably by inflicting oneself (moderate) pain, for example biting or pinching oneself. More subtly, simply imagining an affectively charged situation is enough to induce measurable changes in various physiological systems and such imagery tasks underlie an important part of the empirical data supporting the link between these systems and affective processes.

Interestingly, there is also a large body of literature suggesting that deception itself induces affective changes and measurable activation in bodily systems. Usually, deception research aims at finding some telltale, a response pattern that would betray untruthful answers. Conceivably, careful measurement of several behavioral and physiological variables could enable observers to sort out the different causes underlying a person's behavior, separating the original "genuine" response and the deceptive behavior trying to hide it. But even if it was possible, findings about physiological correlates of deceptive behavior preclude any simplistic assumption about the sensitivity of different type of measurement to lies and conscious attempts at managing one's response.

The upshot of all this research is that simply recording an electrocardiogram or skin conductance does not automatically protect against extraneous influence of the experimental situation on the measurement outcome, be it through unconscious or automatic demand characteristics effects or through willful deception. If uncooperative participants are really a concern, it is absolutely necessary to focus on specific indices (e.g. amplitude of skin conductance responses as opposed to skin conductance in general) and provide a theoretical

rationale and empirical evidence of their relationship with particular processes rather than vague intuitions about the "objectivity" of physiological recordings.

Thinking about inference in psychophysiology is also very relevant to these issues. Cacioppo & Tassinary (1990) provide an historical overview of inference problems within psychophysiology and a model of the different types of relationships between physiological signals and psychological events (see also Fairclough, 2009). A fundamental problem is that psychophysiological research by and large failed to find any strong one-to-one relationship between single physiological signals and psychological processes (called "invariants" by Cacioppo and Tassinary). At best, empirical research identified "outcomes" (physiological responses that are caused by a particular psychological process and therefore always accompany it but that can also be produced by other processes) or "markers" (physiological variables associated with a given psychological process but only in a certain context or for certain participants). Even when an association between a psychological process and a physiological variable (e.g. between emotional arousal and phasic changes in skin conductance) is well documented, other processes (e.g. physical exercise, temperature, mental workload) can cause changes in the physiological variable. This type of many-to-one relationships complicates inference back from the observed changes to a specific process and interpretation of the physiological data is contingent on the ability to control or measure potential confounds, a most difficult proposition for complex stimuli like interactive products.

A similar problem arises in the interpretation of neuropsychological measures (which have, incidentally, also been proposed as a measure of affect in design-oriented research; Motte, 2009). With the increasing availability and performance of brain imaging equipment, many studies attempt to localize specific brain areas that are more closely associated with particular tasks or psychological processes. However, even when sound evidence of increased activation of a given region of the brain during a task exists, it does not mean that there is a one-to-one mapping between activity in this area and the processes engaged by the task. So-called "reverse inference" from the brain imaging data back to the psychological process also requires that no other independent process causes similar patterns of activity. Using a database of neuroimaging results and looking at the example of the famous association between language and Broca's area, Poldrack (2006) shows that this condition is often not met. More than results on "significant" differences of activation between conditions in experiments manipulating a single psychological variable, brain measures require the kind of evidence laid out by Poldrack (2006) to be useful at all.

Examination of the causes of variations in observed measures and potential extraneous variables threatening measurement validity can

also be applied to techniques based on facial expression. As already noted in chapter 2, some types of psychophysiological measures (specifically facial electromyography, a technique that is very similar to electrocardiography in its principle but is used to measure neural control of facial muscles) are really indices of facial behavior and should be considered very differently from signals controlled by the autonomous nervous system (including measures of the cardiovascular system and skin conductance). The neural systems controlling these muscles are also largely separate from the structures regulating bodily arousal and, from a causal perspective, facial electromyography should simply be considered a measure of (expressive) facial behavior.

Unlike autonomic physiology, voluntary control of facial behavior is well documented (Gosselin, Perron & Beaupré, 2010; Rinn, 1984), and there are even less *a priori* reasons to assume that it is immune to the effect of demand characteristics. Indeed, Fridlund & Cacioppo (1986) consider facial electromyography to be more sensitive to demand characteristics than other psychophysiological techniques precisely for this reason, a point largely lost on the applied literature on the measurement of emotion (e.g. Poels & Dewitte, 2006; Motte, 2009). There is however some evidence that facial behavior and self-reported attitudes are not equally sensitive to another threat to validity, namely social desirable responding in prejudice research. Interest for this type of automatically controlled (often called implicit) measures in this field stems from the fact that prejudice is strongly frowned upon in many societies, prompting people who harbor some preferences against a prejudiced group to hide it or even to develop two distinct sets of attitudes (one explicit and conscious when openly discussing the issue and one implicit and unconscious that sometimes manifests itself in behavior). For example, Vanman, Saltz, Nathan & Warren (2004) devised a rather complex procedure that allowed them to measure both self-reported attitudes (friendliness ratings) toward Black and White peoples, facial electromyography in responses to pictures of Black and White people and actual choice in a recruiting tasks in which participants had to choose between three prospective students, based on applications adorned with random pictures of Black and White people[4]. They found that differences in electromyographic activation were related to the final choice of applicant whereas friendliness ratings were not. This means that facial activity seemed essentially immune to social desirable responding and attempts from participants to manage their responses to look good, and could therefore be more useful in predicting affect-related behavior in situations involving socially sensitive issues.

This does not mean however that facial expression is a direct

4 They also used the Implicit Association Test, a common measure of implicit attitudes, with some participants.

reflection of a person's affective state, without interference from any other psychosocial process. Chapter 2 described some theoretical challenges to the notion that facial behavior actually expresses emotion but even the author attributing the greatest role to emotions in facial behavior (namely Paul Ekman) does postulate that two factors drive facial behavior, one of them the innate, stereotypical facial programs constitutive of any affective response, the other being socially and culturally-determined display rules (see chapter 2, section 2.3.5). From a validity standpoint, display rules are an extraneous variable complicating the causal link between affective state and facial expression, threatening the validity of facial behavior observation as a measure of emotions. Audience effects (see e.g. Fridlund, 1997) also belong to the psychosocial factors causing changes on the face that are not solely related to affective processes. In the substantive literature, the influence of these confounding variables is undisputed; the real issue is how they can be accounted for. Still, discussion of emotion measurement based on facial expression analysis in the applied literature (e.g. in affective computing) largely ignores the issue. While nearly all available systems are based directly or indirectly on Ekman's ideas and typology of emotion, their developers and users disregard the logical consequences of his own two-factor model of emotions. They retain the notion of a fixed set of stereotypical basic facial patterns mechanically expressing the current state of the individual but disregard the fact that these basic expressions programs are not the only cause of observable facial behavior.

In fact, a sizable body of research in emotion psychology and facial expression research focuses on the morphological differences between genuine spontaneous affective expressions and controlled or deceptive facial behavior. Several characteristics (e.g. involvement of extra muscle in smiling, dynamics, timing, or symmetry) have been suggested to discriminate between expressions caused by an affective program and expressions caused by display rules or voluntary control. Unfortunately, none of these characteristics have been integrated in current measurement procedures (computer-based automatic analysis of pictures of the face, facial electromyography) so that even if one accepts the most favorable hypotheses from this literature, actual measures of facial expressions cannot claim to be free of the threats to validity discussed until now.

Experimenter expectancy is another potential threat to validity that should be mentioned in a discussion of causal mechanisms and potential threats to validity related to affect measures. Unlike demand characteristics or social desirability, experimenter effects are rarely if ever discussed in relation to emotion measurement in design-related research. It is however a major threat to validity and could have important implications on user experience evaluation practice. Conceptually, experimenter effects can be thought of

as a demand characteristic, an irrelevant variable inconspicuously affecting the measures collected in a test or experiment. In his famous monograph on the subject, Rosenthal (1976) distinguishes several types of experimenter effects. The most straightforward are simple observer errors, differences in interpretation or even intentional errors. These problems affect all sciences (cf. the "personal equation" in astronomy) and are enough to lead different researchers to reach different conclusions about the same phenomenon. Beyond that, behavioral research is also vulnerable to more complex effects resulting from the interaction between experimenters and their research participants. This idea is very similar to Orne's notion of the scientific experiment as a social situation. In this case, researchers are not only biasing the results by observation or interpretation errors but also involuntarily influencing the behavior they observe itself, again running the risk of being unable to replicate each other's results. Rosenthal further distinguishes between several types of interaction between experimenters and their participants including for example biosocial and psychosocial attributes (e.g. gender or personality of the moderator) and situational attributes.

One of the most intriguing types of experimenter effects is however the effect of the experimenter's own orientation towards the outcome of the research. The implications of this experimenter expectancy effect for design-oriented research and product evaluation practice are clear. If the behavior elicited from test participants depends on the researcher's expectation about the results, the outcome of a product test will also depend on the researcher's own attitude towards the design being tested. A new product or a design change could even appear to be an improvement when tested by its promoter and perform worse than existing products when tested by someone bent on killing the project.

In fact, concern about this type of effects underlies double blinding in clinical studies. Properly managing this type of studies involves considerable cost and effort but it has become routine in biomedical fields. It is therefore somewhat surprising that virtually no research seems to be available on the influence of personal variables (experimenter effects and participants' awareness) on product evaluation. Empirical research should establish whether experimenter expectancy effects do also influence perceived usability and user experience and, most importantly, determine the magnitude of these effects for if expectancy effects are markedly smaller than the typical difference of interest between products they need not be a concern for practitioners.

On a more theoretical level, it is interesting to note that while the bulk of the research on experimenter expectancy is based on subjective judgment studies with human participants, Rosenthal reports findings of similar effects in animal studies or response time measures. This is

another example of a psychosocial process threatening measurement validity beyond ostensibly subjective self-report and calling into question any strict separation between "objective" and "subjective" measures.

8.4. Conclusion

The discussion of potential threats to validity contained in this chapter highlighted the complex determinants of measurement outcomes for all major families of emotion measures. Overall, psychophysiological signals and facial expression data are just as complex as self-report ratings and equally sensitive to the top-down psychosocial processes (demand characteristics, social desirability) that are often contrasted with genuine affect. Without a clear causal model of the processes affecting measures, psychological inferences rest on shaky ground. Practitioners interested in the measurement of affect in design and other related fields (human-computer interaction, advertising and consumer psychology) need to attend more carefully to the substantive literature on emotion psychology and to make choices informed by the evidence on the mechanisms underlying variation in the various components of emotion, beyond the simplistic distinction between "objective" and "subjective" measures.

The problem is further compounded by the lack of coherence between these components, observed both in fundamental research with film clips and applied research with computer software. The weakness of these empirical correlations raises a number of practical questions for the evaluation of user experience. It is necessary to define which facet of the users' affective response should be targeted and what processes influence any potential "measure of emotion". Most importantly for design research and actual product tests, it is important to make sure that the components used to measure users' response align with the experiential goals of the design.

In this respect, the most important component of emotion for design research (as opposed to fundamental or clinical research) is often the subjective experience of emotion itself. As far as the person experiencing an emotion is concerned, the phenomenal experience is the emotion and the notions of "pleasure" or "design for experience" refer to the subjective feelings of the users. Simply defining subjective experience away by equating "real" emotion and bodily arousal provides no insight in what creates feelings of pleasure or frustration.

Designers are also likely to be interested in the behavioral consequences of emotions as they can direct the way we interact with the world around us and contribute to our choices and decisions. In this context, other components become somewhat secondary and are only relevant to designers to the extent that they enable them to

predict or shape subjective experience or behavior.

Other components of emotion, especially (facial) expression and physiological activation are generally less useful from a design perspective. The patterns of physiological activation associated with affective processes are obviously interesting in themselves for psychophysiology and neuroscience, but their role in user experience research needs to be considered in the light of the low correlation between emotion components. In most cases, physiological changes or facial expressions are interesting as measures only inasmuch as they can inform us on the subjective feelings of the user. For example, obtaining specific patterns of bodily arousal independently of the broader user experience will seldom be the objective of design practice and research – it could however still be valuable for health applications. These issues need to be weighted carefully in any approach to the assessment of user experience.

9. Conclusion

This thesis presented an approach to moment-to-moment measurement of affect and a series of experiments on emotional experience during short sequences of interaction with products.

In chapter 4, various emotion questionnaires from psychology and design research were shown to be sensitive to differences in experience during interaction with products, both across and within product categories. The results from these studies also documented a level of convergence between different self-report instruments, including different emotion self-report questionnaires and other user experience assessment tools. Comparison with previous research also supported a sensitivity of emotion measures to the task, confirming that actual use and passive observation produced different experiences and that the differences in emotion measures were caused by the interaction with the products and not solely by their appearance.

Chapter 5 extended these results to moment-to-moment ratings of the valence of experience during the interaction itself. Video-supported retrospective self-report (self-confrontation) was shown to be sensitive to product differences and to give insights into the temporal dynamics of the interaction. A study with personal navigation devices also looked at the relationship between these moment-to-moment ratings and overall impression of the product, illustrating the type of research that can be conducted using the method described in this thesis.

Chapter 6 presented the development of the "emotion slider", a device designed to make self-report of emotional valence as intuitive as possible using principles from tangible design and the embodiment of emotion. A series of experiments with pictures established the congruence between the movement necessary to operate the slider and specific emotions, i.e. that affectively charged stimuli preferentially facilitate some behaviors. Asking participants to report the pleasantness of pictures through other, incongruent, movements produces a small but measurable increase in the misclassifications (positive pictures classified as negative or vice versa). Additionally, many participants spontaneously use the amplitude of the movement to express further nuances in the degree of positive or negative valence of each picture despite the fact that neither the instructions nor the feedback given during the experiment explicitly demanded it.

While these studies provided some encouraging data on various aspects of the tools used to measure experience, they also raise a

number of questions. The self-confrontation procedure in particular involves some additional assumptions compared to regular concurrent self-report. While the video is there to help the participants remember their experience and the procedure seems able to give insights into the dynamics of the interaction, validity evidence is indirect, based on correlations with questionnaires and outcome measures. It is relatively easy to add additional measures at the end of a product test or examine correlations between these measures but practically impossible to collect several concurrent series of moment-to-moment self-reports that could be analyzed in a traditional construct validity framework. Direct evidence would directly address the meaning and causes of the moment-to-moment ratings themselves.

For example, further research could determine how much the data collected reflect the (remembered) experience during the use of the product and the role of subsequent elaboration and interpretation by the participants. One way to test the influence of memory on ratings collected during self-confrontation would be to ask *other* respondents to guess what the experience of test participants might have been based on the video recorded during the test ("crossed confrontation") or to vary the delay between the test and the self-confrontation.

The unidimensionality of the measure is another important aspect of the approach that has not been evaluated empirically in the present work. Chapter 4 presented a theoretical and practical rationale for choosing valence as the target dimension but this is of course only one aspect of emotion. The possibility to track other states during self-confrontation or at least to use a bidimensional measure including both valence and arousal should be investigated.

Empirical results also revealed huge individual differences not only in the experience itself but also in the way it was reported using the emotion slider. This was expected but the magnitude of these differences should certainly give pause to researchers in the field. Further thinking on how to deal with these differences and how to articulate different levels of analysis (within-person idiographic accounts and between-persons nomothetic formulations) is clearly needed.

The relationships between the various components of emotion and other experiences should also be investigated further. The low empirical correlations between these components should prompt researchers to think more carefully about what they mean with "experience" or "emotion" and consider whether their definitions and their measures really align with the goals and needs of designers and other practitioners before making claims about the practical relevance of their work. Empirical research should also explore the potential for closer association between specific components of emotion and key behaviors in person-product interaction. If a specific family of measures were found to be better at predicting the way people select,

buy or use products, it would have a particular relevance to design.

More generally, research on the effect products (or indeed other kind of artifacts or stimuli) have on their users raises some specific issues that seem to be largely ignored in the literature. Specifically, for a measurement procedure to be said to reveal anything about a product or the experience it elicits, that product has to play a role in the causal chain leading to the outcome of the measurement. Many inappropriately applied techniques imported from individual differences research (from correlations and reliability coefficients all the way to confirmatory factor analysis and some structural equation models) do not take the various sources of variance in product tests into account and confuse the effects of different designs at the intra-individual level with inter-individual differences. Chapter 7 described the issue and some potential solutions with respect to measurement reliability but the exact same fundamental problem also needs to be solved for validity assessment, both for traditional questionnaires on product appearance, satisfaction, usability, etc. and for moment-to-moment measures of affective responses through self-confrontation, physiological recording or observation of facial behavior.

Further work is needed to sensitize researchers to the issue, and identify and spread techniques to deal with it (e.g. generalizability theory, multi-level factor analysis, etc.) Empirical studies should then investigate how important the differences really are in practice.

The role of emotion dynamics in the formation of the final impression of the product and the overall experience of an interaction sequence also has important implications for design, for example switching the emphasis from the first impression or the average level of pleasure or frustration to the peak and end experiences. Future studies with the approach described in this thesis could help extend these findings to other applications including interaction with software or computer games, service experience or museums. Additionally, bigger samples of participants and products are needed to confirm the peak-end hypothesis and apply more sophisticated analysis techniques that would better use the structure of the self-confrontation data (e.g. time series analysis).

Another way in which moment-to-moment data on the dynamics of emotion could inform design practice is by integrating it directly in the design process, especially in the earlier phases of the process. This can be achieved either by formulating specific recommendations based on the results of a product test (as usually done after usability tests) or simply feeding the data back to the designers (see Desmet, Porcelijn & Van Dijk, 2007 for an example of this approach). Empirical research should compare these approaches and evaluate whether moment-to-moment data are useful at all in the design of interactive products.

A related but even more fundamental question is whether measurement and quantification of (some aspects of) emotional

experience are appropriate at all in the context of product design. It is my conviction that thinking about this question should be informed by an intimate understanding of the main research paradigms down to the nitty-gritty details of concrete methods and not by the kind of casual philosophizing that is often offered as justification for broad theoretical choices. The usefulness of quantification itself and the general realist outlook were therefore understood as useful assumptions and neither as absolute truth or as problems to be tackled within the scope of this thesis.

However, some external arguments for the relevance of quantitative and nomothetic thinking in design-oriented research can be mustered based on the social context of much design activity. Indeed, the existence of design as a distinct profession seems intimately linked to the taylorist structure of industry. Whereas craftsmen traditionally designed and produced small series of objects they could use themselves or even adapt them for individual users, designers define the shape and properties of artifacts that will be mass-produced by other people or even copied identically by machines. There is a fundamental mismatch between the design of widely distributed mass-produced objects to meet the needs and wishes of a range of potential users and research approaches that profess to produce highly specific context-dependent knowledge.

Still, this does not resolve the question or establish that emotions (or some of their attributes like valence) are quantities that can be measured. Quantitative research on user experience seems implicitly based on the view, popular in psychology, that any assignment of numbers to objects or events following any specified set of operations constitutes quantification. It is however by no means self-evident that all attributes are actually quantitative and the quantitative nature of any particular attribute must be established empirically to support its measurement (Michell, 1999). This is a thorny question that still seems insufficiently explored both in the general literature on emotion and in application-oriented measurement efforts.

A common reason to perform user research of any kind – quantitative or not – is that it is often difficult for designers to empathize with the future users of the product they are designing and to predict their needs and preferences based solely on their own personal experience. From this perspective, measures of emotion should act as a bridge between designers and users, recording and aggregating their responses and subjectivity in understandable and actionable insights for the designers. Interestingly, much work on emotion in applied fields starts with a conceptualization of emotion inherited from psychological research or implicitly based on the target population intuitive understanding of the phenomenon. Looking at emotion from the perspective of designers and other consumers

of user experience measures could provide another perspective on emotion in design and help present the results from user research in a way that is relevant and useful to practitioners.

10. References

Aaker, D.A., Stayman, D.M., & Hagerty, M.R. (1986). Warmth in advertising: Measurement, impact, and sequence effects. *Journal of Consumer Research*, 12, 365-381.

Abelson, R.P. (1995). *Statistics as Principled Argument.* Hillsdale, NJ: Lawrence Erlbaum.

Alexopoulos, T., & Ric, F. (2007). The evaluation-behavior link: Direct and beyond valence. *Journal of Experimental Social Psychology*, 43 (6), 1010-1016.

Baayen, R.H., Davidson, D.J., & Bates, D.M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language,* 59 (4), 390-412.

Baguley, T. (2009). Standardized or simple effect size: What should be reported? British *Journal of Psychology,* 100 (3), 603-617.

Bamford, S., & Ward, R. (2008). Predispositions to Approach and Avoid Are Contextually Sensitive and Goal Dependent. *Emotion*, 8 (2), 174-183.

Barrett, L.F. (2006). Valence is a basic building block of emotional life. *Journal of Research in Personality*, 40 (1), 35-55.

Barrett, L.F., & Russell, J.A. (1999). The structure of current affect: Controversies and emerging consensus. *Current Directions in Psychological Science,* 8 (1), 10-14.

Barsalou, L.W. (2009). Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364 (1521), 1281-1289.

Bartlett, M.S., Hager, J.C., Ekman, P., & Sejnowski, T.J. (1999). Measuring facial expressions by computer image analysis. *Psychophysiology*, 36 (2), 253-263.

Bartlett, M.S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., & Movellan, J. (2006). Fully automatic facial action recognition in spontaneous behavior. *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition, FGR 2006 (pp. 223-230).*

Bassili, J.N. (1978). Facial motion in the perception of faces and of emotional expression. *Journal of Experimental Psychology: Human Perception and Performance*, 4 (3), 373-379.

Bassili, J.N. (1979). Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face. *Journal of Personality and Social Psychology*, 37 (11), 2049-2058.

Berntson, G.G., Bigger, Jr., J.T., Eckberg, D.L., Grossman, P., Kaufmann, P.G., Malik, M., Nagaraja, H.N., Porges, S.W., Saul, J.P., Stone, P.H., & Van der Molen, M.W. (1997). Heart rate variability: Origins, methods, and interpretive caveats. *Psychophysiology*, 34, 623-648.

Berridge, K., & Winkielman, P. (2003). What is an unconscious emotion? (The case for unconscious "liking"). *Cognition and Emotion,* 17 (2), 181-211.

Biocca, F., David, P., & West, M. (1994). Continuous Response Measurement (CRM): A computerized Tool for Research on the Cognitive Processing of Communication Messages. In A. Lang (Ed.), *Measuring Psychological Response to Media* (pp. 15-64). Hillsdale, N.J.: Erlbaum Associates.

Blackwell, A.F., Fitzmaurice, G., Holmquist, L.E., Ishii, H. & Ullmer, B. (2007). *Tangible User Interfaces in Context and Theory.* Workshop at the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2007), April 28–May 3, 2007, San Jose, CA.

Blythe, M. A., Overbeeke, K., Monk, A. F., & Wright, P. C. (2003). *Funology: from usability to enjoyment*. Boston: Kluwer Academic Publishers.

Bonanno, G.A., & Keltner, D. (2004). The coherence of emotion systems: Comparing "on-line" measures of appraisal and facial expressions, and self-report. *Cognition and Emotion,* 18 (3), 431-444

Borsboom, D., Mellenbergh, G.J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review,* 111 (4), 1061-1071.

Boucsein, W. (1992). *Electrodermal Activity*. New York: Plenum Press.

Boucsein, W., & Backs, R.W. (2000). Engineering Psychophysiology as a Discipline: Historical and Theoretical Aspects. In R. W. Backs & W. Boucsein (Eds.) *Engineering Psychophysiology* (pp. 3-29), Mahwah, NJ: Lawrence Erlbaum.

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.

Bradley, M.M., & Lang, P.J. (2000). Affective reactions to acoustic stimuli. *Psychophysiology*, 37 (2), 204-215.

Bradley, M.M., & Lang, P.J. (2007). The International Affective Picture System (IAPS) in the Study of Emotion and Attention. In J.A. Coan, & J.J.B Allen (Eds.), *Handbook of Emotion Elicitation and Assessment* (pp. 29-46). Oxford: Oxford University Press.

Bradley, M.M., Miccoli, L., Escrig, M.A., & Lang, P.J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, 45 (4), 602-607.

Brennan, R.L. (2001). *Generalizability Theory.* New York: Springer.

Brittin, R.V., & Duke, R.A. (1997). Continuous versus summative evaluations of musical intensity: A comparison of two methods for measuring overall effect. *Journal of Research in Music Education*, 45, 245-258.

Broekens, J., Pronker, A., & Neuteboom, M. (2010). *Real Time Labelling of Affect in Music Using AffectButton*. Paper presented at the ACM Multimedia Conference, October 25-29, 2010, Firenze, Italy.

Brooke, J. (1996). SUS – A Quick and dirty usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & A. L. McClelland (Eds.), *Usability Evaluation in Industry*. London: Taylor and Francis.

Cacioppo, J.T., Berntson, G.G., Larsen, J.T., Poehlmann, K.M, & Ito, T.A. (2000). The psychophysiology of emotion. In R. Lewis, & J.M. Haviland-Jones (Eds.), *The handbook of emotion, 2nd edition* (pp. 173-191). New York: Guilford Press.

Cacioppo, J.T., & Petty, R.E. (1979). Attitudes and cognitive response: An electrophysiological approach. *Journal of Personality and Social Psychology*, 37 (12), 2181-2199.

Cacioppo, J.T., Petty, R.E., Losch, M.E., & Kim, H.S. (1986). Electromyographic Activity Over Facial Muscle Regions Can Differentiate the Valence and Intensity of Affective Reactions. *Journal of Personality and Social Psychology*, 50 (2), 260-268.

Cacioppo, J.T., & Tassinary, L.G. (1990). Inferring Psychological Significance from Physiological Signals. *American Psychologist,* 45 (1), 16-28.

Cacioppo, J. T., Tassinary, L. G., & Fridlund, A. F. (1990). The skeletomotor system. In J. T. Cacioppo and L. G. Tassinary (Eds.), *Principles of psychophysiology: Physical, social, and inferential elements* (pp. 325-384). New York: Cambridge University Press.

Cahour, B., Salembier, P., Brassac, C., Bouraoui, J.L., Pachoud, B., Vermersch, P., & Zouinar, M. (2005). *Methodologies for evaluating the affective experience of a mediated interaction*. Paper presented at the Workshop on Innovative Approaches to Evaluating Affective Interfaces, ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2005), April 2-7, 2005, Portland, OR.

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin,* 56 (2), 81-105.

Chen, M., & Bargh, J.A. (1999). Consequences of automatic evaluation: Immediate behavioral predispositions to approach or avoid the stimulus. *Personality and Social Psychology Bulletin*, 25 (2), 215-224.

Clark, H.H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior,* 12 (4), 335-359.

Cleary, T.A., & Linn, R.L. (1969). Error of measurement and the power of a statistical test. *British Journal of Mathematical and Statistical Psychology,* 22, 49-55.

Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences* (Rev. ed.). New York: Academic Press.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist,* 49 (12), 997-1003.

Cohn, J.F., & Ekman, P. (2005). Measuring Facial Action. In J. Harrigan, R. Rosenthal, & K.R. Scherer (Eds.), *New Handbook of Methods in Nonverbal Behavior Research* (pp. 9-64). Oxford: Oxford University Press.

Cohn, J.F., & Kanade, T. (2007). Use of Automated Facial Image Analysis for Measurement of Emotion Expression. In J.A. Coan, & J.J.B Allen (Eds.), *Handbook of Emotion Elicitation and Assessment* (pp. 222-238). Oxford: Oxford University Press.

Cohn, J.F., Kanade, T., Moriyama, T., Ambadar, Z., Xiao, J., Gao, J., & Imamura, H. (2001). *A Comparative Study of Alternative FACS Coding Algorithms*. Technical Report CMU-RI-TR-02-06.

Cohn, J.F., Zlochower, A.J., Lien, J., & Kanade, T. (1999). Automated face analysis by feature point tracking has high concurrent validity with manual FACS coding. *Psychophysiology*, 36 (1), 35-43.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M. (2000). *'feeltrace': an instrument for recording perceived emotion in real time*. Paper presented at the ISCA Workshop on Speech and Emotion, September 5–7, 2000, Newcastle, United Kingdom.

Crawford, J.R., & Henry, J.D. (2004). The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample. *British Journal of Clinical Psychology*, 43 (3), 245-265.

Crocker, P. R. E. (1997). A Confirmatory Factor Analysis of the Positive Affect Negative Affect Schedule (PANAS) With a Youth Sport Sample. *Journal of Sport and Exercise Psychology*, 19, 91-97.

De Winter, J.C.F., Dodou, D., & Wieringa, P.A. (2009). Exploratory factor analysis with small sample sizes. *Multivariate Behavioral Research,* 44 (2), 147-181.

Den Uyl, M., & van Kuilenburg, H. (2005). *The FaceReader: Online facial expression recognition*. Paper presented at the 5th International Conference on Methods and Techniques in Behavioral Research (Measuring Behavior 2005), August, 30–September 2, 2005, Wageningen, The Netherlands.

Desmet, P.M.A. (2002). *Designing Emotions*. PhD Thesis, Delft University of Technology, The Netherlands.

Desmet, P.M.A. (2004). Measuring Emotions. Development and application of an instrument to measure emotional responses to products. In M.A. Blythe, A.F. Monk, K. Overbeeke, & P.C. Wright (Eds.), *Funology: from usability to enjoyment*. Dordrecht, The Netherlands: Kluwer.

Desmet, P.M.A., & Dijkhuis, E. (2003). *A Wheelchair can be Fun: A Case of Emotion-driven Design*. Paper presented at the International Conference on Designing Pleasurable Products and Interfaces (DPPI 2003), June 23–26, 2003, Pittsburgh, PA.

Desmet, P.M.A., & Hekkert, P. (2002). The basis of product emotions. In W. Green & P. Jordan (Eds.), *Pleasure with Products, Beyond Usability* (pp. 60-68). London: Taylor & Francis.

Desmet, P.M.A., & Hekkert, P. (2007). Framework of product experience. *International Journal of Design,* 1 (1), 57-66.

Desmet, P.M.A., Hekkert, P., & Hillen, M.G. (2004). *Values and emotions; an empirical investigation in the relationship between emotional responses to products and human values.* Paper presented at Techné: Design Wisdom, 5th European Academy of Design Conference, April 2003, Barcelona, Spain.

Desmet, P.M.A, Hekkert, P., & Jacobs, J.J. (2000). When a car makes you smile: Development and applications of an instrument to measure product emotions. *Advances in Consumer Research*, 27, 111-117.

Desmet P.M.A., Porcelijn, R., & van Dijk, M. (2007). Emotional design; Application of a research based design approach. *Journal of Knowledge, Technology & Policy*, 20 (3), 141-155.

Desmet, P.M.A., & Schifferstein, R. (2010). *Holistic & dynamic experience – first explorations* (unpublished report). Delft, the Netherlands.

Dimberg, U. (1988). Facial electromyography and the experience of emotion. *Journal of Psychophysiology*, 2 (4), 277-282.

Dimberg, U., & Karlsson, B. (1997). Facial reactions to different emotionally relevant stimuli. *Scandinavian Journal of Psychology*, 38 (4), 297-303.

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

Dubé, L., & Morgan, M. S. (1996). Trend Effects and Gender Differences in Retrospective Judgments of Consumption Emotions. *The Journal of Consumer Research*, 23 (2), 156-162.

Duckworth, K.L., Bargh, J.A., Garcia, M., & Chaiken, S. (2002). The automatic evaluation of novel stimuli. *Psychological Science*, 13 (6), 513-519.

Duke, R.A., & Colprit, E.J. (2001). Summarizing Listener Perceptions Over Time. *Journal of Research in Music Education*, 49 (4), 330-342.

Eder, A.B., & Rothermund, K. (2008). When Do Motor Behaviors (Mis)Match Affective Stimuli? An Evaluative Coding View of Approach and Avoidance Reactions. *Journal of Experimental Psychology: General*, 137 (2), 262-281.

Eibl-Eibsfeldt, I. (1997). *Die Biologie des Menschenlischen Verhaltens. Grundriß der Humanethologie. 3rd edition.* München: Piper.

Ekman, P. (1992). Facial expressions of emotion: New findings, new questions. *Psychological Science,* 3 (1), 34-38.

Ekman, P. (1993). Facial expression and emotion. *American Psychologist*, 48 (4), 384-392.

Ekman, P. (1999). Facial Expressions. In T. Dalgleish, & M. Power, *Handbook of Cognition and Emotion* (pp. 45-60). New York: John Wiley & Sons Ltd.

Ekman, P., & Friesen, W.V. (1969). The repertoire of nonverbal behavior: categories, origins, usage, and coding. *Semiotica*, 1, 49-98.

Ekman, P. (1972). Universals and cultural differences in facial expressions of emotion. In J. Cole (Ed.), *Nebraska symposium on motivation 1971* (pp. 207-283). Lincoln: University of Nebraska Press.

Ekman, P. (1994). Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. *Psychological Bulletin*, 115, 268-287.

Ekman, P. (1999). Facial Expressions. In T. Dalgleish, & M. Power (Eds.), *Handbook of Cognition and Emotion* (pp. 45-60). New York: John Wiley & Sons Ltd.

Ekman, P., & Friesen, W.V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17 (2), 124-129.

Ekman, P., & Friesen, W.V. (1978). *Facial Action Coding System: A technique for the measurement of facial movement*. Palo Alto, CA: Consulting Psychologists Press.

Ekman, P., Friesen, W.V., & Hager, J.C. (2002). *Facial Action Coding System.* (electronic form)

Ekman, P., Sorenson, E.R., & Friesen, W.V. (1969). Pan-cultural elements in facial displays of emotion. *Science*, 164 (3875), 86-88.

Elfenbein, H.A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin*, 128 (2), 203-235.

Fabrigar, L.R., MacCallum, R.C., Wegener, D.T., & Strahan, E.J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods,* 4 (3), 272-299.

Fairclough, S.H. (2009). Fundamentals of physiological computing. *Interacting with Computers,* 21 (1-2), 133-145.

Feldt, L.S., & Brennan, R.L. (1989). Reliability. In R.L Linn (Ed.), *Education Measurement* (3rd ed.) (pp. 105-146). New York: Macmillan.

Fontaine, J.R.J., Scherer, K.R., Roesch, E.B., & Ellsworth, P.C. (2007). The world of emotions is not two-dimensional. *Psychological Science,* 18 (12), 1050-1057.

Fowles, D.C., Christie, M.J., Edelberg, R., Grings, W.W., Lykken, D.T., & Venables, P.H. (1981). Publication Recommendations for Electrodermal Measurements. *Psychophysiology*, 18 (3), 232-239.

Fredrickson, B.L. (2001). The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. *American Psychologist*, 56 (3), 218-226.

Fredrickson, B.L., & Kahneman, D. (1993). Duration Neglect in Retrospective Evaluations of Affective Episodes. *Journal of Personality and Social Psychology*, 65 (1), 45-55.

Fridlund, A.J. (1991). Sociality of Solitary Smiling: Potentiation by an Implicit Audience. *Journal of Personality and Social Psychology*, 60 (2), 229-240.

Fridlund, A.J. (1997). The new ethology of human facial expressions. In J. A. Russell & J. M. Fernández-Dols (Eds.), *The psychology of facial expression: Studies in emotion and social interaction* (pp. 103-129). New York: Cambridge University Press.

Fridlund, A.J., & Cacioppo, J.T. (1986). Guidelines for Human Electromyographic Research. *Psychophysiology*, 23 (5), 567-589.

Frijda, N.H., & Tcherkassof, A. (1997). Facial expressions as modes of action readiness. In J. A. Russell & J. M. Fernández-Dols (Eds.), *The psychology of facial expression: Studies in emotion and social interaction* (pp. 78-102). New York: Cambridge University Press.

Geringer, J.M., Madsen, C.K., & Gregory, D. (2004). A fifteen-year history of the Continuous Response Digital Interface: Issues relating to validity and reliability. *Bulletin of the Council for Research in Music Education*, (160), 1-10.

Gosling, S.D., Rentfrow, P.J., & Swann Jr., W.B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, 37 (6), 504-528.

Gosselin, P., Perron, M., & Beaupré, M. (2010). The Voluntary Control of Facial Action Units in Adults. *Emotion*, 10 (2), 266-271.

Gotlib, I.H., & Meyer, J.P. (1986). Factor Analysis of the Multiple Affect Adjective Check List. A Separation of Positive and Negative Affect. *Journal of Personality and Social Psychology*, 50 (6), 1161-1165.

Gottman, J.M., & Levenson, R.W. (1985). A Valid Procedure for Obtaining Self-Report of Affect in Marital Interaction. *Journal of Consulting and Clinical Psychology,* 53 (2), 151-160.

Gross, J.J., & Levenson, R.W. (1997). Hiding feelings: The acute effects of inhibiting negative and positive emotion. *Journal of Abnormal Psychology*, 106 (1), 95-103.

Hassenzahl, M. (2004). The interplay of beauty, goodness, and usability in interactive products. *Human-Computer Interaction*, 19 (4), 319-349.

Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In G. Szwillus, & J. Ziegler (Eds.), *Mensch & Computer 2003: Interaktion in Bewegung* (pp. 187-196). Stuttgart: B.G. Teubner.

Havlena, W. J., & Holbrook, M. B. (1986). The Varieties of Consumption Experience: Comparing Two Typologies of Emotion in Consumer Behavior. *The Journal of Consumer Research*, 13 (3), 394-404.

Hazlett, R. L. (2003). *Measurement of User Frustration: A Biologic Approach.* Paper presented at the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2003), April 5–10, 2003, Fort Lauderdale, FA.

Hess, E.H., & Polt, J.M. (1960). Pupil size as related to interest value of visual stimuli. *Science*, 132 (3423), 349-350.

Hewig, J., Hagemann, D., Seifert, J., Gollwitzer, M., Naumann, E., & Bartussek, D. (2005). A revised film set for the induction of basic emotions. *Cognition and Emotion*, 19 (7), 1095-1109.

Hogan, T.P., Benjamin, A., & Brezinski, K.L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement,* 60 (4), 523-531.

Holbrook, M.B., & Westwood, R.A. (1989). The Role of Emotion in Advertising Revisited: Testing a Typology of Emotional Responses. In P. Cafferata & A.M. Tybout (Eds.), *Cognitive and Affective Responses to Advertising* (pp. 353-371). Lexington, MA: Lexington Books.

Huang, M.-H. (1997). Is negative affect in advertising general or specific? A comparison of three functional forms. *Psychology and Marketing*, 14 (3), 223-240.

Hugdahl, K. (1995). *Psychophysiology. The Mind-Body Perspective*. Cambridge, MA: Harvard University Press.

Izard, C.E. (1971). *The face of emotion*. New York: Appleton-Century-Crofts.

Izard, C.E. (1994). Innate and universal facial expressions: Evidence from developmental and cross-cultural research. *Psychological Bulletin*, 115 (2), 288-299.

Izard, C.E., & Dougherty, L.M. (1982). Two complementary systems for measuring facial expressions in infants and children. In C.E. Izard (Ed.), *Measuring emotions in infants and children, Volume 1*. Cambridge: Cambridge University Press.

Jenkins, S., Brown, R., & Rutterford, N. (2009). Comparing Thermographic, EEG, and Subjective Measures of Affective Experience During Simulated Product Interactions. *International Journal of Design*, 3 (2), 53-65.

Jennings, J.R., Berg, W.K., Hutcheson, J.S., Obrist, P., & Porges, S. (1981). Publication Guidelines for Heart Rate Studies in Man. *Psychophysiology*, 18 (3), 226-231.

Jensen, R. (1999). *The Dream Society*. New York: McGraw-Hill.

Jordan, P.W. (2000). *Designing Pleasurable Products*. London: Taylor & Francis.

Kane, M.T. (2001). Current concerns in validity theory. *Journal of Educational Measurement,* 38 (4), 319-342.

Karapanos, E. (2010). *Quantifying Diversity in User Experience*. PhD Thesis, Eindhoven University of Technology.

Karapanos, E., Zimmerman, J., Forlizzi, J., & Martens, J.-B. (2010). Measuring the dynamics of remembered experience over time. *Interacting with Computers*, 22 (5), 328-335.

Katsikitis, M., Pilowsky, I., & Innes, J.M. (1990). The quantification of smiling using a microcomputer-based approach. *Journal of Nonverbal Behavior*, 14 (1), 3-17.

Kellerman, H., & Plutchik, R. (1968). Emotion-trait interrelations and the measurement of personality. *Psychological Reports*, 23 (3), 1107-1114.

King, S.C., & Meiselman, H.L. (2010). Development of a method to measure consumer emotions associated with foods. *Food Quality and Preference*, 21 (2), 168-177.

Kreibig, S.D. (2010). Autonomic nervous system activity in emotion: A review. *Biological Psychology*, 84 (3), 394-421.

Kring, A.M., & Sloan, D.M. (2007). The Facial Expression Coding System (FACES): Development, Validation, and Utility. *Psychological Assessment*, 19 (2), 210-224.

Krone, A., Hamborg, K.-C., & Gediga, G. (2002). Zur emotionalen reaktion bei fehlern in der mensch-computer-interaktion. *Zeitschrift fur Arbeits- und Organisationspsychologie,* 46 (4), 185-200.

Lance, C.E., Butts, M.M., & Michels, L.C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods*, 9 (2), 202-220.

Lang, P.J., Greenwald, M.K., Bradley, M.M., & Hamm, A.O. (1993). Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology*, 30 (3), 261-273.

Larsen, J.T., Berntson, G.G., Poehlmann, K.M., Ito, T.A., & Cacioppo, J.T. (2008). The psychophysiology of emotion. In R. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), *The handbook of emotions* (3rd ed.) (pp. 180-195). New York: Guilford.

Larsen, J.T., Norris, C.J., & Cacioppo, J.T. (2003). Effects of positive and negative affect on electromyographic activity over *zygomaticus major* and *corrugator supercilii*. *Psychophysiology*, 40 (5), 776-785.

Larsen, J.T., Norris, C.J., McGraw, A.P., Hawkley, L.C., & Cacioppo, J.T. (2009). The evaluative space grid: A single-item measure of positivity and negativity. *Cognition and Emotion,* 23 (3), 453-480.

Laurans, G. (2009). [Pre-test PANAS ratings from various experiments]. Unpublished raw data.

Law, E., Roto, V., Hassenzahl, M., Vermeeren, A., & Kort, J. (2009). *Understanding, Scoping and Defining User Experience: A Survey Approach*. Paper presented at the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2009), April 4–9, 2009, Boston, MA.

Lee, K.P. & Jeong, S.H. (2006). *Development of Tool for Video-Debriefing for Understanding Emotion over Usability*. Paper presented at the 5th Conference on Design & Emotion, September 27–29, 2006, Gothenburg, Sweden.

Levenson, R.W. (2003). Blood, Sweat, and Fears: The Autonomic Architecture of Emotion. *Annals of the New York Academy of Sciences,* 1000, 348-366.

Lim, S. S. (2002). The Self-Confrontation Interview: Enhancing our Understanding of Human Factors in Web-based Interaction. *Journal of Electronic Commerce,* 3 (3), 162-173.

Liu, Y., & Salvendy, G. (2009). Effects of measurement errors on psychometric measurements in ergonomics studies: Implications for correlations, ANOVA, linear regression, factor analysis, and linear discriminant analysis. *Ergonomics,* 52 (5), 499-511.

Loftus, G.R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science,* 5 (6), 161-171.

Lorr, M. (1989). Models and Methods for Measurement of Mood. In R. Plutchik, & H. Kellerman (Eds.), *The Measurement of Emotion* (pp. 37-53). San Diego, CA: Academic Press.

Lorr, M., & Wunderlich, R.A. (1988). A semantic differential mood scale. *Journal of Clinical Psychology*, 44 (1), 33-36.

Ludden, G.D.S. (2008). *Sensory incongruity and surprise in product design.* PhD Thesis, Delft University of Technology.

Ludden, G.D.S., Schifferstein, H.N.J., & Hekkert, P. (2006). *Sensory Incongruity, Comparing Vision to Touch, Audition, and Olfaction.* Paper presented at the 5th Conference on Design & Emotion, September 27–29, 2006, Gothenburg, Sweden.

Lychner, J. (1998). An empirical study concerning terminology relating to aesthetic response to music. *Journal of Research in Music Education*, 46 (2), 303-319.

MacCallum, R.C., Widaman, K.F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods,* 4 (1), 84-99.

MacCallum, R.C., Widaman, K.F., Preacher, K.J., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research,* 36 (4), 611-637.

Mackinnon, A., Jorm, A.F., Christensen, H., Korten, A.E., Jacomb, P.A., & Rodgers, B. (1999). A short form of the Positive and Negative Affect Schedule: Evaluation of factorial validity and invariance across demographic variables in a community sample. *Personality and Individual Differences*, 27 (3), 405-416.

Mahlke, S., Minge, M., & Thüring, M. (2006). *Measuring Multiple Components of Emotions in Interactive Contexts*. Paper presented at the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2006), April 22–27, 2006, Montréal, Canada.

Mahlke, S., & Thüring, M. (2007). *Studying Antecedents of Emotional Experiences in Interactive Contexts.* Paper presented at the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2007), April 28–May 3, 2007, San Jose, CA.

Mandryk, R.L., & Atkins, M.S. (2007). A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. *International Journal of Human Computer Studies*, 65 (4), 329-347.

Markman, A.B., & Brendl, C.M. (2005). Constraining theories of embodied cognition. *Psychological Science,* 16 (1), 6-10.

Marsh, A.A., Ambady, N., & Kleck, R.E. (2005). The effects of fear and anger facial expressions on approach- and avoidance-related behaviors. *Emotion*, 5 (1), 119-124.

Matsumoto, D., Ekman, P., & Fridlund, A. (1991). Analyzing Nonverbal Behavior. In P.W Dowrick (Ed.), *Practical Guide to Using Video in the Behavioral Sciences* (pp. 153-165). New York: Wiley & Sons.

Matsumoto, D., Keltner, D., Shiota, M.N., O'Sullivan, M., & Frank, M. (2008). Facial Expressions of Emotion. In M. Lewis, J.M. Haviland-Jones, & L. Feldman Barrett (Eds.), *Handbook of Emotions* (3rd ed.) (pp. 211-234). New York: Guilford.

Matsumoto, D., & Willingham, B. (2006). The thrill of victory and the agony of defeat: Spontaneous expressions of medal winners of the 2004 Athens olympic games. Journal of Personality and Social Psychology, 91 (3), 568-581.

Mauss, I.B., McCarter, L., Levenson, R.W., Wilhelm, F.H., & Gross, J.J. (2005). The tie that binds? Coherence among emotion experience, behavior, and physiology. *Emotion,* 5 (2), 175-190.

McDonagh, D., Hekkert, P. van Erp, J., & Gyi, D. (Eds.) (2003). *Design and Emotion, Episode III: The experience of everyday things*. London: Taylor & Francis.

Mehrabian, A. (1995). Framework for a comprehensive description and measurement of emotional states. *Genetic, social, and general psychology monographs,* 121 (3), 339-361.

Mehrabian, A. (1996). Pleasure-Arousal-Dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology,* 14 (4), 261-292.

Mehrabian, A., & Russell, J.A. (1974). *An Approach to Environmental Psychology.* Cambridge, Massachussetts: The MIT Press.

Meier, B.P., & Robinson, M.D. (2004). Why the Sunny Side Is Up: Associations Between Affect and Vertical Position. *Psychological Science*, 15 (4), 243-247.

Merla, A., & Romani, G.L. (2007). *Thermal signatures of emotional arousal: A functional infrared imaging study.* Paper presented at the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. August 22–26, 2007, Lyon, France.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist,* 50 (9), 741-749.

Michell, J. (1999). *Measurement in psychology critical history of a methodological concept*. New York: Cambridge University Press.

Mikels, J.A., Fredrickson, B.L., Larkin, G.R., Lindberg, C.M., Maglio, S.J., & Reuter-Lorenz, P.A. (2005). Emotional category data on images from the international affective picture system. *Behavior Research Methods*, 37 (4), 626-630.

Milgram, S. (1974). *Obedience to Authority: An Experimental View*. New York: Harper & Row.

Mooradian, T.A., & Olver, J.M. (1997). "I can't get no satisfaction:" The impact of personality and emotion on postpurchase processes. *Psychology and Marketing*, 14 (4), 379-393.

Motte, D. (2009). *Using Brain Imaging to Measure Emotional Response to Product Appearance.* Paper presented at the International Conference on Designing Pleasurable Products and Interfaces (DPPI 2009), October 13–16, 2009, Compiègne, France.

Mundfrom, D.J., Shaw, D.G., & Ke, T.L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing,* 5 (2), 159-168.

Nagel, F., Kopiez, R., Grewe, O., & Altenmüller, E. (2007). EMuJoy: Software for continuous measurement of perceived emotions in music. *Behavior Research Methods*, 39 (2), 283-290.

Niedenthal, P.M. (2007). Embodying emotion. *Science*, 316 (5827), 1002-1005.

Norman, D.W. (2004). *Emotional Design: Why we love (or hate) everyday things.* New York: Basic Books.

Nowlis, V. (1965). Research with the Mood Adjective Check List. In S.S. Tomkins & C.E. Izard (Eds). *Affect, Cognition and Personality* (pp. 352-389). New York: Springer-Verlag.

Nunnally, J.C. (1967). *Psychometric Theory.* New York: McGraw Hill.

Orne, M.T. (1962). On the Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and Their Implications. *American Psychologist,* 17, 776-783.

Overall, J.E., & Woodward, J.A. (1975). Unreliability of difference scores: A paradox for measurement of change. *Psychological Bulletin,* 82 (1), 85-86.

Pantic, M. (2009). Machine analysis of facial behaviour: Naturalistic and dynamic behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364 (1535), 3505-3513.

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

Papillo, J. F., & Shapiro, D. (1990). The cardiovascular system. In J. T. Cacioppo & L. G. Tassinary (Eds.) *Principles of psychophysiology: Physical, social, and inferential elements* (pp. 456-512). New York: Cambridge University Press.

Parkinson, B. (2005). Do facial movements express emotions or communicate motives? *Personality and Social Psychology Review,* 9 (4), 278-311.

Partala, T., & Surakka, V. (2003). Pupil size variation as an indication of affective processing. *International Journal of Human Computer Studies*, 59 (1-2), 185-198.

Peeters, F.P.M.L., Ponds, R.W.H.M., & Vermeeren, M.T.G. (1996). Affectiviteit en zelfbeoordeling van depressie en angst. *Tijdschrift voor Psychiatrie*, 38 (3), 240-250.

Picard, R.W. (2010). Affective Computing: From laughter to IEEE. *IEEE Transactions on Affective Computing,* 1 (1), 11-17.

Pilowsky, I., & Katsikitis, M. (1994). The classification of facial emotions: A computer-based taxonomic approach. *Journal of Affective Disorders*, 30 (1), 61-71.

Plutchik, R. (1966). Multiple rating scales for the measurement of affective states. *Journal of Clinical Psychology*, 22 (4), 423-425.

Plutchik, R. (1980). *Emotion: A psychoevolutionary synthesis.* New York: Harper & Row.

Poels, K., & Dewitte, S. (2006). How to capture the heart? Reviewing 20 years of emotion measurement in advertising. *Journal of Advertising Research*, 46 (1), 18-37.

Poldrack, R.A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences,* 10 (2), 59-63.

Power, M.J. (2006). The structure of emotion: An empirical comparison of six models. *Cognition and Emotion*, 20 (5), 694-713.

Preacher, K.J., & MacCallum, R.C. (2002). Exploratory factor analysis in behavior genetics research: Factor recovery with small sample sizes. *Behavior Genetics,* 32 (2), 153-161.

Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics,* 2, 13-32.

Puri, C., Olson, L., Pavlidis, I., Levine, J., Starren, J. (2005). *StressCam: Non-contact Measurement of Users' Emotional States through Thermal Imaging.* Paper presented at the ACM Conference on Human Factors in Computing Systems (CHI 2005), April 2–7, 2005, Portland, OR.

Raaijmakers, J.G.W., Schrijnemakers, J.M.C., & Gremmen, F. (1999). How to Deal with "The Language-as-Fixed-Effect Fallacy": Common Misconceptions and Alternative Solutions. *Journal of Memory and Language,* 41 (3), 416-426.

Ravaja, N., Turpeinen, M., Saari, T., Puttonen, S., & Keltikangas-Järvinen, L. (2008). The Psychophysiology of James Bond: Phasic Emotional Responses to Violent Video Game Events. *Emotion*, 8 (1), 114-120.

Redelmeier, D.A., & Kahneman, D. (1996). Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain*, 66 (1), 3-8.

Reeve, C.L., Highhouse, S., & Brooks, M.E. (2006). A closer look at reactions to realistic recruitment messages. *International Journal of Selection and Assessment*, 14 (1), 1-15.

Revelle, W. (2009). *psych: Procedures for Psychological, Psychometric, and Personality Research*. R package version 1.0-85. http://CRAN.R-project.org/package=psych

Richins, M. L. (1997). Measuring Emotions in the Consumption Experience. *The Journal of Consumer Research*, 24 (2), 127-146.

Rinn, W.E. (1984). The neuropsychology of facial expression: A review of the neurological and psychological mechanisms for producing facial expressions. *Psychological Bulletin,* 95 (1), 52-77.

Rosenthal, R. (1976). *Experimenter effects in behavioral research*. New York, NY: Appleton-Century-Crofts.

Rotteveel, M., & Phaf, R.H. (2004). Automatic affective evaluation does not automatically predispose for arm flexion and extension. *Emotion*, 4 (2), 156-172.

Ruef, A.M., & Levenson, R.W. (2007). Continuous Measurement of Emotion. The Affect Rating Dial. In J.A. Coan & J.J.B. Allen (2007), *Handbook of Emotion Elicitation and Assessment* (pp. 286-297). Oxford: Oxford University Press.

Russell, J. A. (1980). A Circumplex Model of Affect. *Journal of Personality and Social Psychology,* 39 (6), 1161-1178.

Russell, J.A. (1994). Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin*, 115 (1), 102-141.

Russell, J.A. (1995). Facial expressions of emotion: What lies beyond minimal universality? *Psychological Bulletin*, 118 (3), 379-391.

Russell, J.A. (2003). Core Affect and the Psychological Construction of Emotion. *Psychological Review,* 110 (1), 145-172.

Russell, J. A., Bachorowski, J.-A., & Fernández-Dols, J.-M. (2003). Facial and Vocal Expressions of Emotion. *Annual Review of Psychology*, 54, 329-349.

Russell, J. A., & Fernández-Dols, J.-M. (1997). What does a facial expression mean? In J. A. Russell & J. M. Fernández-Dols (Eds.), *The psychology of facial expression: Studies in emotion and social interaction* (pp. 3-30). New York, NY: Cambridge University Press.

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

Russell, J.A., Weiss, A., & Mendelsohn, G.A. (1989). Affect Grid: A Single-Item Scale of Pleasure and Arousal. *Journal of Personality and Social Psychology*, 57 (3), 493-502.

Russo, B. (2010). *Shoes, Cars, and Other Love Stories: Investigating the Experience of Love for Products*. PhD Thesis, Delft University of Technology.

Sato, W., Fujimura, T., & Suzuki, N. (2008). Enhanced facial EMG activity in response to dynamic facial expressions. *International Journal of Psychophysiology*, 70 (1), 70-74.

Scherer, K.R. (1984). Emotion as a multicomponent process: A model and some cross-cultural data. *Review of Personality and Social Psychology*, 5, 37-63.

Scherer, K.R. (2005). What are emotions? and how can they be measured? *Social Science Information*, 44 (4), 695-729.

Scherer, K.R., & Grandjean, D. (2008). Facial expressions allow inference of both emotions and their components. *Cognition and Emotion*, 22 (5), 789-801.

Scherer, K.R., Schorr, A., & Johnstone, T. (Eds.) (2001). *Appraisal processes in emotion.* Oxford: Oxford University Press.

Schubert, E. (1999). Measuring emotion continuously: Validity and reliability of the two-dimensional emotion-space. *Australian Journal of Psychology,* 51 (3), 154-165.

Schubert, E. (2001). Continuous Measurement of Self-Report Emotional Response to Music. In P. Juslin & J. Sloboda (Eds.), *Music and Emotion: Theory and Research* (pp. 393-414). Oxford, UK: Oxford University Press.

Schwartz, G.E., Fair, P.L., Salt, P., Mandel, M.R., & Klerman, G.L. (1976). Facial muscle patterning to affective imagery in depressed and nondepressed subjects. *Science,* 192 (4238), 489-491.

Seibt, B., Neumann, R., Nussinson, R., & Strack, F. (2008). Movement direction or change in distance? Self- and object-related approach-avoidance motions. *Journal of Experimental Social Psychology*, 44 (3), 713-720.

Shaffer, J.P. (1995). Multiple hypothesis testing. *Annual Review of Psychology,* 46 (1), 561-584.

Shapiro, D., Jamner, L.D., Lane, J.D., Light, K.C., Myrtek, M., Sawada, Y., & Steptoe, A. (1996). *Psychophysiology*, 33, 1-12.

Shavelson, R.J., & Webb, N.N. (1991). *Generalizability Theory. A Primer.* Newbury Park, CA: SAGE.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika,* 74 (1), 107-120.

Solarz, A.K. (1960). Latency of Instrumental Responses as a Function of Compatibility with the meaning of Eliciting Verbal Signs. *Journal of Experimental Psychology*, 59 (4), 239-245.

Stayman, D.M., & Aaker. D.A. (1993). Continuous Measurement of Self-Report of Emotional Response. *Psychology and Marketing*, 10 (3), 199-214.

Sutcliffe, J.P. (1958). Error of measurement and the sensitivity of a test of significance. *Psychometrika,* 23 (1), 9-17.

Tellegen, A., Watson, D., & Clark, L.A. (1999). On the dimensional and hierarchical structure of affect. *Psychological Science,* 10 (4), 297-303.

Thayer, J.F., & Sinclair, R.C. (1987). Psychological distress: A hierarchical factor model of the multiple affect adjective check list (MAACL). *Journal of Psychopathology and Behavioral Assessment*, 9 (2), 229-233.

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools,* 44 (5), 423-432.

Thompson, E.R. (2007). Development and validation of an internationally reliable short-form of the Positive and Negative Affect Schedule (PANAS). *Journal of Cross-Cultural Psychology*, 38 (2), 227-242.

Tran, V. (2004). *The Influence of Emotions on Decision-Making Processes in Management Teams.* Unpublished doctoral dissertation, Université de Genève, Geneva, Switzerland.

Tukey, J. W. (1991). The Philosophy of Multiple Comparisons. *Statistical Science,* 6 (1), 100-116.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, 58 (1), 6-20.

Vacha-Haase, T., Kogan, L.R., & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement,* 60 (4), 509-522.

Valstar, M., & Pantic, M. (2006). *Fully automatic facial action unit detection and temporal analysis.* Paper presented at the Conference on Computer Vision and Pattern Recognition (CVPR 2006), June 17–22, 2006, New York.

Van Dantzig, S., Pecher, D., & Zwaan, R.A. (2008). Approach and avoidance as action effects. *Quarterly Journal of Experimental Psychology*, 61 (9), 1298-1306.

Van Dantzig, S, Zeelenberg, R., & Pecher, D. (2009). Unconstraining theories of embodied cognition. *Journal of Experimental Social Psychology,* 45 (2), 345-351.

Van Kuilenburg, H., Wiering, M., & den Uyl, M. (2005). *A model based method for automatic facial expression recognition.* Paper presented at the 16th European Conference on Machine Learning (ECML 2005), October 3–7, 2005, Porto, Portugal.

Vanden Abeele, P., & MacLachlan, D. L. (1994). Process Tracing of Emotional Responses to TV Ads: Revisiting the Warmth Monitor. *The Journal of Consumer Research*, 20 (4), 586-600.

Vanman, E.J., Saltz, J.L., Nathan, L.R., & Warren, J.A. (2004). Racial discrimination by low-prejudiced whites - Facial movements as implicit measures of attitudes related to behavior. *Psychological Science,* 15 (11), 711-714.

Visch, V.T., & Goudbeek, M.B. (2009*). Emotion attribution to basic parametric static and dynamic stimuli.* Paper presented at the 3rd International Conference on Affective Computing and Intelligent Interaction (ACII 2009), September 10–12, 2009, Amsterdam, The Netherlands.

Wagenmakers, E.-J. & Brown, S.D. (2007). On the linear relationship between the mean and the standard deviation of a response time distribution. *Psychological Review*, 114 (3), 830-841.

Wallbott, H.G. (1998). Bodily expression of emotion. *European Journal of Social Psychology*, 28 (6), 879-896.

Wang, Y.J., & Minor, M.S. (2008). Validity, reliability, and applicability of psychophysiological techniques in marketing research. *Psychology and Marketing*, 25 (2), 197-232.

Ward, R.D., & Marsden, P.M. (2003). Physiological responses to different WEB page designs. *International Journal of Human Computer Studies*, 59 (1-2), 199-212.

Watson, D., & Clark, L.A. (1994). *Manual for the Positive and Negative Affect Schedule - Expanded Form (PANAS-X).* University of Iowa.

Watson, D., Clark, L.A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology*, 54 (6), 1063-1070.

Watson, D., Wiese, D., Vaidya, J., & Tellegen, A. (1999). The two general activation systems of affect: Structural evolutionary considerations, and psychobiological evidence. *Journal of Personality and Social Psychology*, 76 (5), 820-838.

Westbrook, R. A., & Oliver, R. L. (1991). The Dimensionality of Consumption Emotion Patterns and Consumer Satisfaction. *The Journal of Consumer Research*, 18 (1), 84-91.

Westerman, S.J., Sutherland, E.J., Robinson, L., Powell, H., & Tuck, G. (2007). A multi-method approach to the assessment of web page designs. *2nd International Conference on Affective Computing and Intelligent Interaction, ACII 2007, September 12-14, 2007.*

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. New York: Springer.

Wilcox, R.R. (1996). Confidence intervals for the slope of a regression line when the error term has nonconstant variance. *Computational Statistics and Data Analysis*, 22 (1), 89-98.

Wilcox, R.R. (2003). *Applying Contemporary Statistical Techniques*. San Diego, CA: Academic Press.

Wilcox, R.R. (2005). *Introduction to Robust Estimation and Hypothesis Testing. 2nd edition.* Burlington, MA: Elsevier.

Williams, R.H., & Zimmerman, D.W. (1989). Statistical power analysis and reliability of measurement. *Journal of General Psychology,* 116, 359-369.

Williams, R.H., Zimmerman, D.W., & Zumbo, B.D. (1995). Impact of measurement error on statistical power: Review of an old paradox. *Journal of Experimental Education,* 63, 363-370.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist,* 54, 594-604.

Woltman Elpers, J.L.C.M., Wedel, M., & Pieters, R.G.M. (2003). Why Do Consumers Stop Viewing Television Commercials? Two Experiments on the Influence of Moment-to-Moment Entertainment and Information Value. *Journal of Marketing Research,* 40 (4), 437-453.

Wright, S.P. (1992). Adjusted P-values for simultaneous inference. *Biometrics,* 48 (4), 1005-1013.

Yik, M.S.M., Russell, J.A., & Barrett, L.F. (1999). Structure of self-reported current affect: Integration and beyond. *Journal of Personality and Social Psychology,* 77 (3), 600-619.

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.

Zeitlin, D.M., & Westwood, R.A. (1986). Measuring emotional response. *Journal of Advertising Research*, 26 (5), 34-44.

Zeng, Z., Pantic, M., Roisman, G.I., & Huang, T.S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 31 (1), 39-58.

Zuckerman, M., & Lubin, B. (1990). A Useful Measure for State Affects. *Current Contents,* 31, 24.

Zuckerman, M., Lubin, B., & Rinck, C.M. (1983). Construction of new scales for the multiple affect adjective check list. *Journal of Behavioral Assessment*, 5 (2), 119-129.

Zuckerman, M., Lubin, B., Rinck, C.M., Soliday, S.M., Albott, W.L., & Carlson, K. (1986). Discriminant validity of the Multiple Affect Adjective Check List - revised. *Journal of Psychopathology and Behavioral Assessment*, 8 (2), 119-128.

Zumbo, B.D. (2007). Validity: Foundational Issues and Statistical Methodology. In C.R. Rao & S. Sinharay (Eds.), *Handbook of Statistics: Vol. 26. Psychometrics* (pp. 45-79). New York: Elsevier.

Zwick, W.R., & Velicer, W.F. (1986). Comparison of Five Rules for Determining the Number of Components to Retain. *Psychological Bulletin,* 99 (3), 432-442.

Appendix A. PrEmo factor analysis

Several studies in the thesis (personal navigation device study in chapters 3 and 5, coffee machine and alarm clock study in chapter 3) use different variants of the PrEmo questionnaire (Desmet, 2002) to assess users' emotional experience after using a product. While this questionnaire is intended primarily as a measure of distinct categorical emotions like dissatisfaction, disgust or joy, PrEmo ratings tend to exhibit moderate to strong correlations.

In fact, most measures of distinct emotions have often been shown to share a sizable amount of common variance and it is likely that higher order factors like positive and negative activation and valence can be extracted from PrEmo data (see chapter 2 for relevant references and more details on current models of emotion). While information about pleasantness tends to be less suggestive to designers than specific emotions (Desmet, 2002), deriving a measure of valence from PrEmo data can be useful for a number of reasons, for example to obtain more reliable measures, compare PrEmo data with other measures or perform an overall evaluation of the difference in experience between two products.

This appendix presents a factor analysis conducted to evaluate the dimensionality of PrEmo, using data from the personal navigation device study (see chapter 3). This analysis was performed on data pooled across the different products used in the study[1]. Parallel analysis and scree test (figure A.1) both suggested that only one factor should be retained[2]. The single factor represented 45% of the variance in the data.

1 See appendix B and chapter 7 for some limitations of this type of 'disaggregation'.

2 Parallel analysis was conducted using the *fa.parallel* function in William Revelle's *psych* package for R (Revelle, 2009). See also appendix B for more detail on parallel analysis and factor retention decisions.

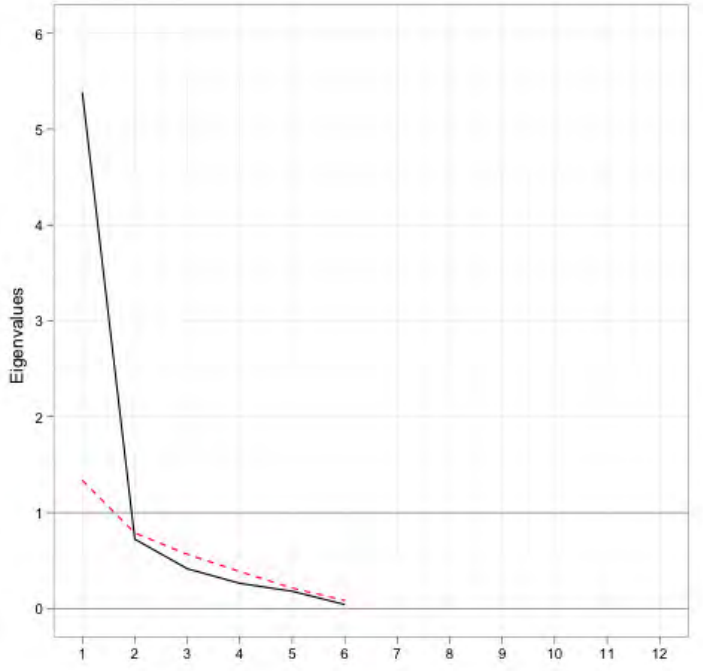


Figure A.1. *Scree plot for PrEmo data in personal navigation study. The black line represents the eigenvalues of the correlation matrix obtained from the actual data; the red line corresponds to simulated data in parallel analysis*[3].

The factor matrix indicates that the structure of the scale largely conforms to the expectations, with a clear bipolar structure anchored by emotions of opposite valence. All positive emotions are strongly correlated with this valence factor but negative emotions have somewhat lower loadings (table A.1). Several negative emotions (contempt, unpleasant surprise, boredom) have relatively small communalities (under .2).

3 Since the analysis presented here is a factor analysis, the scree plot and parallel analysis are based on the *reduced* correlation matrix (i.e. a correlation matrix with estimates of the communalities in the diagonal; see Fabrigar, MacCallum, Wegener & Strahan, 1999).

Table A.1. *Factor matrix resulting from a principal axis factor analysis of PrEmo ratings of personal navigation devices*[4].

Item	
contempt	-.31
dissatisfaction	-.85
unpleasant surprise	-.38
disgust	-.68
boredom	-.42
sadness	-.60
admiration	.80
satisfaction	.81
pleasant surprise	.72
desire	.82
fascination	.61
joy	.74

4 Principal axis factoring was performed using the *factor.pa* function in William Revelle's *psych* package for R.

Appendix B. Component analysis of product meaning questionnaire

This appendix presents an analysis of the structure of the product meaning questionnaire used in the study on personal navigation devices presented in chapter 3 (section 3.2).

Since the study used a between-subject design, participants are nested within the main conditions (i.e. the personal navigation device used) and correlations computed across the whole data set confuse participant-related variation and product-related variation. The sample size is also very small compared to traditional guidelines for this type of analyses (but see appendix C for a discussion of this problem). For all these reasons, the results presented here are only offered as very exploratory findings.

The significant differences between the products' mean scores on the various scales defined through this analysis do however suggest that the correlations really do reflect product-related variation, at least partly, and the relationship with the other questionnaires used in the study (see below) are also encouraging.

Oblique rotations suggest that the various factors in these ratings are far from independent. However, since results from factor and component analyses with different oblique (Promax) and orthogonal rotations were broadly similar (i.e. the same set of items related to each factor), only the somewhat antiquated but much more common truncated principal component analysis with Varimax rotation will be discussed here.

Kaiser's traditional eigenvalue over 1 criterion suggested retaining five components but parallel analysis supported a three-component solution (figure B.1)[1]. Since the three-component structure was also

1 In spite of being the default setting in SPSS/PASW, the "eigenvalue over 1" factor retention criterion overstates the actual number of factors or components in many situations (Lance, Butts & Michels, 2006; Zwick & Velicer, 1986) and its use has been consistently discouraged in the recent literature on factor analysis (Fabrigar, MacCallum, Wegener & Strahan, 1999; Preacher & MacCallum, 2003). Parallel analysis is often recommended as an alternative. The general principle is to generate random matrices with the same aggregate characteristics (number of variables, sample size, communalities)

more interpretable and corresponded to previous results obtained with the same questionnaire (Desmet & Schifferstein, 2010), only this solution will be discussed further (see table B.1 for the rotated component matrix). The three rotated components represented 30%, 15% and 10% of the total variance in the data.

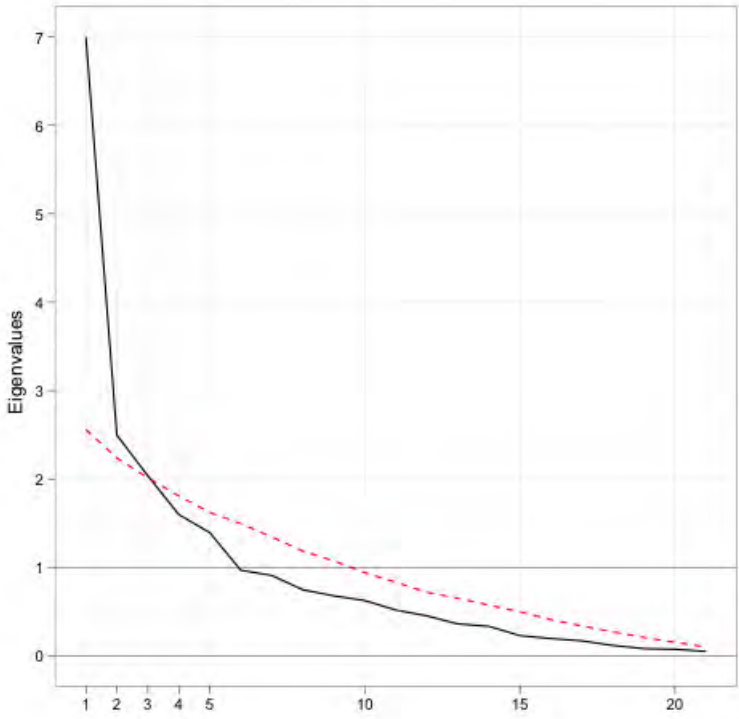


Figure B.1. *Scree plot for PrEmo data in personal navigation study. The black line represents the eigenvalues of the correlation matrix obtained from the actual data; the red line corresponds to simulated data in parallel analysis.*

The results of the principal component analysis are not very different from previous results with the same questionnaire, especially for the first two components. The first rotated component reflects the function and usefulness of the device and is associated with words like "helpful", "handy" or "smart". The second rotated component seems

than the data being analyzed and to compute the average eigenvalues for these simulated matrices. These eigenvalues represent the results that would be expected if the data were pure noise with no particular structure. The number of eigenvalues in the original data matrix exceeding these simulated eigenvalues then indicates the number of meaningful factors/components that can be extracted. Parallel analysis was conducted using the *fa.parallel* function in William Revelle's *psych* package for R (Revelle, 2009).

related to more intangible properties and to the style of the product. The pattern of loadings on the third component was less clear, with very few items specifically associated with it and many cross-loadings. It was therefore dropped from all further analyses.

Table B.1. Component matrix resulting from a truncated principal component analysis followed by Varimax rotation on meaning ratings of personal navigation devices[2]. Correlations in bold correspond to the items used to compute scale scores.

Item			
helpful	**.90**	.19	.00
handy	**.88**	.09	-.07
stimulating	**.83**	.08	-.10
smart	**.81**	.28	-.02
clear	**.75**	.28	.05
reliable	**.74**	.05	-.02
balanced	**.72**	-.06	.24
abundant	**-.66**	.04	.21
interesting	.64	.43	-.09
restless	-.45	-.25	.23
cheap	-.07	**-.70**	-.03
distinctive	.02	**.70**	-.33
playful	.28	**.67**	.12
oldfashioned	.11	**-.63**	.16
stylish	.24	.59	.40
tough	.24	.54	.15
intimidating	-.18	-.23	.10
abstract	-.11	-.09	.78
businesslike	.42	.03	.63
authentic	.44	.19	-.55
headstrong	-.32	.35	.52

Based on these results, two simple summative scales were devised. The scores for the first scale, called helpfulness, were computed by adding item ratings for "helpful", "handy", "stimulating", "smart", "clear", "reliable", "balanced", and "abundant". "Interesting", "businesslike", and "authentic" were not retained because of cross-loadings with other components. "Restless" was also dropped because of the relatively small correlation with the component.

The scores for the second scale, called distinctiveness, were obtained by adding the ratings for "cheap", "distinctive", "playful", and "old-fashioned". "Stylish" and "tough" were not included because

2 This analysis was performed using PASW 18.0.

of the somewhat smaller correlations with the component and, for the former, because of a large cross-loading. For both scales, items with negative loadings were inverted before summing them.

Scores on these two scales were compared to the results of other questionnaires used in the same study (table B.2). These correlations reveal a great deal of convergence between helpfulness, AttrakDiff's pragmatic quality, and the System Usability Scale. Distinctiveness is more specifically correlated to AttrakDiff's hedonic quality. Correlations between these two groups of scales are lower, but not negligible.

Table B.2. *Correlation between selected questionnaires in the personal navigation device study*

	1	**2**	**3**	**4**	**5**	**6**
1. System Usability Sc.	1					
2. Pragmatic Quality	**.93**	1				
3. Helpfulness	**.91**	**93**	1			
4. Hedonic Qual. Stimul.	.41	.30	.31	1		
5. Distinctiveness	.44	.27	.26	**.74**	1	
6. Emotion	**.88**	**.83**	**.87**	.56	.46	1

Appendix C. Note on sample sizes in factor and component analyses

Appendix A and B present some component and factor analyses for questionnaire data used in chapter 3. While the outcome matched theoretical expectations and previous results with the same questionnaires, the sample size in these studies might seem unusually low for this type of analyses.

Studies in personality psychology or educational testing are often conducted on data sets with thousands or tens of thousands of observations, a sample size several orders of magnitude larger than those practiced in design research. Even sample sizes in the hundreds are rare in this field and are usually obtained with relatively lightweight data collection techniques (i.e. mail or internet surveys rather than actual product tests and self-confrontation).

This state of affairs is in stark contrast with traditional guidelines from the psychological literature on factor analysis, which typically recommend 100 to 300 participants as absolute minimum and a participants-to-variables ratio between 3 to 1 and 20 to 1 (Mundfrom, Shaw & Ke, 2005). Factor analysis should however not be prematurely ruled out as an analysis strategy for product ratings as recent simulation studies have shown that, under certain conditions, good results are possible with much smaller sample sizes (De Winter, Dodou & Wieringa, 2009; Mundfrom et al.; Preacher & MacCallum, 2002).

In fact, several factors other than the sample size, including the level of communality and the number of variables per factor (overdetermination) affect the quality of the results (De Winter, Dodou & Wieringa, 2009; Mundfrom et al., 2005; Preacher & MacCallum, 2002). For example, in the most favorable conditions (high communalities, a single factor and 5 to 8 variables), Mundfrom et al. found that as few as 11 participants are sufficient to get good results. Conversely, in the most difficult conditions in their simulations (low communalities, more than 3 factors and 3 variables per factor), 1200 observations are necessary to reach the same level of congruence between the population model and the factor analyses results, making any absolute recommendation or guidelines based solely on the number of variables completely irrelevant to judge the sample size (in these examples the participants-to-variable ratio of the minimum

sample size range from almost 1 to 1 to a worse case of 1 to 130, well over any published recommendation).

Fortunately, the data presented in appendices A and B have several characteristics (overdetermination, moderate to strong level of communality) that would seem to make factor analyses viable, certainly for PrEmo data. However, it must be noted that simulations are often based on relatively simple cases and many aspects that could complicate the analysis (correlation between factors, non-normal discrete distributions – attenuating or distorting correlations) have not been comprehensively examined in the literature yet. Another difficulty is that beside well-determined factors, real data also typically include nuisance factors and variables with high cross-loadings or low communalities that could threaten the analysis. Additionally, in most studies, the population model is unknown and the assessment of the sample size is based in part on the sample data matrix. For example, the adequateness of the sample size strongly depends on the number of factors in the population (or, equivalently when the number of variables is fixed, to the variables-to-factors ratio) but in the most exploratory studies (e.g. appendix B), the only information available on the number of factors to be extracted results from the analysis of a potentially inadequate sample. Still, factor or component analyses should not be ruled out merely on the basis of irrelevant guidelines or the modest sample size in these studies.

Curriculum vitae

Gaël Laurans was born on the 28th of March 1981 in Saint-Julien-en-Genevois (France). He attended the Louis-Dumont *collège* and obtained a science *baccalauréat* at the Saint-Exupéry *lycée* in Bellegarde-sur-Valserine (France).

From 1998 to 2000 he studied applied computer science at the *Institut universitaire de technologie Nancy-Charlemagne* (University Nancy 2, France) and obtained a *diplôme universitaire de technologie* followed in 2001 by a National Diploma in Computing (with distinction) from the Institute of Technology, Sligo (Ireland).

He then switched to cognitive science and obtained a *licence* (2002) and *maîtrise* (2003) in cognitive science and a *diplôme d'études supérieures spécialisées* in occupational psychology from the University of Metz and the University of Nancy 2 in 2004.

In 2005, he started his PhD research on the measurement of emotion at the faculty of Industrial Design Engineering of Delft's University of Technology.

Summary

This thesis investigated the measurement of emotion during short episodes of interaction between products and their users.

Chapter 2 is a review of the many ways that have been used to measure emotions, organized according to the component of emotion involved: feelings, bodily changes, and facial expression.

Measurement based on bodily changes and facial expression is costly and requires extensive expertise. Still, several physiological measures have been considered in the design-related literature but they often lack specificity. Even if automatic recognition systems have recently become available, applied research based on the observation of facial expression remains extremely rare. Both physiological recording and facial expression recognition could in principle have huge advantages for moment-to-moment assessment of emotion as they provide nearly continuous data without requiring the active participation of the research participants. However, their lack of reliability forces researchers to rely on multiple trials and averaging in analysis, thus precluding simple online measurement.

Self-report, based on conscious feelings, is easier to apply and is the most common way to measure emotions. Self-report measurement instruments based on different models of emotion are available including measures of pleasantness and arousal and measures of discrete emotions like anger or disgust. Several of these questionnaires have been used in a design context, often to assess responses to product appearance or long-term use. Moment-to-moment self-report is also common in fields like advertisement or music research but is typically limited to dimensional models of emotion (measuring pleasantness or arousal).

Chapter 3 is devoted to punctual measures of emotion in person-product interaction. It describes two studies in which participants had to complete different questionnaires right after using a product. The first study compared two questionnaires chosen for their extensive coverage of positive emotions – PrEmo and the Geneva Emotion Wheel – in a test with a coffee machine and an alarm clock. The results show both instruments to be sensitive to differences between products and document a decent level of convergence between the questionnaires.

The second study extended these results to a between-subject experimental design in which each participant only used one of the products tested. It found a variant of PrEmo to be sensitive to

differences between several personal navigation devices and examined the relationships between measures of different aspects of user experience (perceived usability, meaning, feelings).

Chapter 4 is devoted to continuous or moment-to-moment measures of emotion in person-product interaction. It describes the particular challenges facing researchers interested in the dynamics of ongoing emotional changes during the interaction itself. It then sketches an approach developed to tackle this problem, by combining several techniques used in other fields. A key element of this approach is a technique called self-confrontation. It uses video to collect time-bound data about specific events right after the interaction while avoiding interrupting as it unfolds.

Chapter 5 describes two studies conducted with the approach developed in chapter 4. The first study asked participants to report about their experience using two vases, selected to be either frustrating or surprising. The second study collected data about the pleasantness or unpleasantness of a drive using one of several personal navigation devices. The differences between the products were found to be related to specific parts of the routes the participants had to follow. The results also suggest that the peak experience (how bad the experience was at its worse or how good it was at its best) is more important in determining the overall experience than the average experience over the whole test.

Chapter 6 describes the development of a device, the emotion slider, conceived to make moment-to-moment self-report more intuitive following the principles of tangible design. An experiment using pictures as affective stimuli was conducted before using the emotion slider to collect moment-to-moment data about dynamic stimuli. Following some unexpected results, a series of experiments was organized to better understand the properties of the slider. These experiments showed that the link between movement and affect is more complex than initially thought.

Chapter 7 discusses reliability and its impact for applied measurement. It starts with a brief review of key concepts and of the limitations of some common measures of reliability. A numerical example shows that these measures can be misleading when improperly applied to data about transient states like product-related emotions as opposed to individual traits like personality and intelligence. Generalizability theory, a technique that can be used to deal with these issues is introduced through a re-analysis of some the data from chapter 3.

Chapter 8 is devoted to the notion of measurement validity. After a review of the most salient perspectives on validity within psychometrics, the data presented in chapters 3 and 5 are re-evaluated. The chapter also contains a discussion of several conceptual issues regarding the validity of measures derived from different components of emotion.

Samenvatting

Dit proefschrift onderzocht het meten van emotie tijdens korte periodes van interactie tussen producten en hun gebruikers.

Hoofdstuk 2 is een overzicht van de vele manieren die zijn gebruikt om emotie te meten, ingedeeld op basis van het betrokken aspect van emotie: gevoelens, lichamelijke veranderingen en gezichtsuitdrukkingen.

Meten op basis van lichamelijke veranderingen en gezichtsuitdrukkingen is duur en vergt uitgebreide expertise. Toch zijn verscheidene fysiologische maten beproefd in de ontwerpliteratuur. Deze missen echter vaak specificiteit. Toegepast onderzoek op basis van het observeren van gezichtsuitdrukkingen blijft bijzonder zeldzaam, zelfs nu daarvoor recent automatische herkenningssystemen beschikbaar zijn gekomen.

Zowel fysiologische metingen als gezichtsuitdrukkingsherkenning kunnen in principe enorme voordelen bieden voor de beoordeling, van moment tot moment, van emotie aangezien zij een nagenoeg continue datastroom verzorgen waarvoor geen actieve handeling van de deelnemers aan het onderzoek is vereist. Echter, de gebrekkige betrouwbaarheid van deze metingen dwingt onderzoekers om meervoudige tests te gebruiken en te middelen in de analyse, wat eenvoudige online metingen uitsluit.

Zelfrapportage, gebaseerd op bewuste gevoelens, is eenvoudiger toe te passen en is de meest voorkomende manier om emoties te meten. Er zijn meetinstrumenten beschikbaar voor zelfrapportage die gebaseerd zijn op verschillende emotiemodellen waaronder maten voor plezierigheid en opwinding en maten voor discrete emoties zoals boosheid of walging. Enkele van deze vragenlijsten zijn gebruikt in een ontwerpcontext, vaak om reacties te peilen op het uiterlijk van een product of op het gebruik over langere termijn. Zelfrapportage van moment tot moment is ook gebruikelijk op het terrein van het adverteren en in muziekonderzoek maar is dan in de regel beperkt tot dimensionele emotiemodellen (het meten van plezierigheid of opwinding).

Hoofdstuk 3 is gewijd aan het meten van emotie op één of enkele momenten tijdens (een onderbreking in) de interactie tussen mens en product. Het bevat twee onderzoeken waarin deelnemers verschillende vragenlijsten moesten invullen direct na het gebruik van een product.

Het eerste onderzoek vergelijkt twee vragenlijsten die zijn uitgekozen vanwege hun uitgebreide behandeling van positieve emoties - PrEmo

en The Geneva Emotion Wheel – in tests met een koffiezetapparaat en een wekker. De resultaten tonen aan dat beide vragenlijsten in staat zijn om verschillen tussen producten te registreren en laten een behoorlijke convergente validiteit zien tussen de vragenlijsten.

Het tweede onderzoek lag in het verlengde van deze resultaten en had een tussen-subjectenopzet waarin elke deelnemer slechts één van de geteste producten gebruikte. Dit toonde aan dat een variant van PrEmo in staat is verschillen tussen enkele persoonlijke navigatieapparaten te registreren en onderzocht de relatie tussen de gemeten waardes van verschillende aspecten van de gebruikersbeleving (waargenomen gebruiksvriendelijkheid, betekenis, gevoelens).

Hoofdstuk 4 is gewijd aan het continue of van moment tot moment meten van emotie tijdens de interactie tussen mens en product. Het beschrijft de bijbehorende uitdagingen voor onderzoekers die geïnteresseerd zijn in de dynamiek van emotionele veranderingen zoals die zich voordoen tijdens de interactie zelf. Het beschrijft vervolgens een methode om dit probleem aan te pakken die is ontwikkeld door enkele technieken uit andere terreinen te combineren. Een sleutelelement in deze methode is een techniek die zelfconfrontatie heet. Dit behelst het gebruik van video om tijdgebonden data te verzamelen over bepaalde gebeurtenissen onmiddellijk volgend op de interactie en dus zonder de interactie te onderbreken terwijl deze plaatsvindt.

Hoofdstuk 5 beschrijft twee onderzoeken die zijn uitgevoerd met de methode uit hoofdstuk 4. Het eerste onderzoek vroeg deelnemers te rapporteren over hun gebruikersbeleving met twee vazen die waren geselecteerd om frustrerend, respectievelijk verassend te zijn. Het tweede onderzoek verzamelde gegevens over de plezierigheid of onplezierigheid van een rit waarbij gebruik werd gemaakt van één van enkele persoonlijke navigatieapparaten. De verschillen tussen producten bleken gerelateerd te zijn aan bepaalde delen van de routes die de deelnemers moesten volgen. De resultaten suggereren tevens dat de ervaringspieken (hoe slecht de ervaring was op zijn slechtst en hoe goed op zijn best) belangrijker zijn voor het bepalen van de uiteindelijke gebruikerservaring dan de gemiddelde ervaring tijdens de hele test.

Hoofdstuk 6 beschrijft de ontwikkeling van een apparaat, de emotie-schuifknop, die is bedacht om de zelfrapportage van moment tot moment meer intuïtief te maken, geïnspireerd door tangible design principes. Een experiment met plaatjes als affectieve stimuli werd eerst uitgevoerd en daarna is de emotie-schuifknop gebruikt om van moment tot moment data te verzamelen over dynamische stimuli. Na enige onverwachte resultaten is een serie experimenten opgezet om de eigenschappen van de schuifknop beter te begrijpen. Deze experimenten tonen aan dat de koppeling tussen beweging en affect complexer is dan eerst werd gedacht.

Hoofdstuk 7 bediscussieert betrouwbaarheid en de invloed daarvan op toegepast meten. Het begint met een kort overzicht van sleutelbegrippen en van de beperkingen van enkele veelgebruikte maten van betrouwbaarheid. Een numeriek voorbeeld laat dan zien deze maten misleidend kunnen zijn indien onjuist toegepast op data over steeds veranderende toestanden zoals product-gerelateerde emoties (in tegenstelling tot individuele kenmerken zoals persoonlijkheid en intelligentie). Generaliseerbaarheidstheorie, een techniek die gebruikt kan worden om deze problemen op te lossen wordt vervolgens geïntroduceerd door middel van een her-analyse van een deel van de data uit hoofdstuk 7.

Hoofdstuk 8 is gewijd aan de validiteit van meetmethodes. Na een overzicht van de belangrijkste opvattingen over validiteit binnen de psychometrie worden de gegevens uit hoofdstuk 3 en 5 opnieuw geëvalueerd. Dit hoofdstuk bevat tevens een verhandeling over enkele conceptuele problemen aangaande de validiteit van maten die gebaseerd zijn op verschillende componenten van emotie.

Acknowledgments

There are so many people who have helped me over the course of the last six years that I cannot hope not to forget anyone. I must therefore only hope that those I forget will forgive me.

I am grateful to all my colleagues at the faculty of Industrial Design Engineering, especially to Elif Özcan-Vieira for welcoming me when everybody else was away in the summer of 2005, for her graphic design tips and for her general kindness, to Geke Ludden for providing the vases I used in chapter 5, to Anna Fenko and Jeroen Arendsen for the many interesting discussions, to Michel Varkevisser for sharing his knowledge of psychophysiology and always being available when I had a question, and to Cha Joong Kim for serving both as a model and as a photographer on several occasions. I would also like to thank Rob Luxen and Hannah Ottens for their role in the realization of the emotion slider, Marc de Hoogh, Bertus Naagen, Henk Lok, and Arend Harteveld.

I would also like to thank the secretaries of the Industrial design department, Carla Gerbracht, Annemarie Metselaar, Sonja Grinsven-Evers, Monique van Biljouw, Amanda Klumpers-Nieuwpoort, Daphne van der Does, and Ashley Marapin. Without them, nothing would be running.

I am also thankful to the many students, interns, and research assistants whose work contributed to this thesis (David Güiza Caicedo and Marleen van Beuzekom, Max Braams, Maarten Langbroek, and Jorn Ouborg, Lara van der Veen, Ahmet Bektes, Remon de Wijngaert) and to those who volunteered to participate in my experiments.

Finally, I would like to thank Erik Groenenberg for his help with the final layout of the thesis.