A Case-Study of Affect Measurement Tools for Physical User Interface Design

Colin Swindells[1] Karon E. MacLean[1] Kellogg S. Booth[1] Michael Meitner[2]

[1] Department of Computer Science
University of British Columbia
{swindell, maclean, ksbooth}@cs.ubc.ca

[2] Department of Forest Resources Management
University of British Columbia
meitner@interchange.ubc.ca

ABSTRACT

Designers of human-computer interfaces often overlook issues of affect. An example illustrating the importance of affective design is the frustration many of us feel when working with a poorly designed computing device. Redesigning such computing interfaces to induce more pleasant user emotional responses would improve the user's health and productivity. Almost no research has been conducted to explore affective responses in rendered haptic interfaces. In this paper, we describe results and analysis from two user studies as a starting point for future systematic evaluation and design of rendered physical controls. Specifically, we compare and contrast self-report and biometric measurement techniques for two common types of haptic interactions. First, we explore the tactility of real textures such as silk, putty, and acrylic. Second, we explore the kinesthetics of physical control renderings such as friction and inertia. We focus on evaluation methodology, on the premise that good affect evaluation and analysis cycles can be a useful element of the interface designer's tool palette.

CR Categories: H.1.2 [User/Machine Systems]: Human factors, Human information processing, Software psychology; H.5.2 [Information Interfaces and Presentation]: User Interfaces–Haptic I/O; D.2.2 [Software Engineering]: Tools and Techniques–User interfaces

Keywords: affect, emotion, design, user interface, human factors experiment, haptics, manual controls.

1 MOTIVATION

Affect computing refers to computing devices that relates to, arises from, or deliberately influences one's emotions [18]. Furthermore, our affective responses always companion thought [26]. For example, we rarely see a "house"; instead, we see a handsome house, an ugly house, a pretentious house [28]. In terms of computer systems, we see a cool, sleek new computer, hear an upbeat cell phone ring tone, or feel a comfortable stylus. Such affective judgments are believed to be independent of, and temporarily precede most higher-level perceptual and cognitive operations. In other words, affective responses are a 'first level' response to our environment [1]. These 'gut' affective responses then influence higher-level emotional judgements, which are more cognitive. Consequently, higher-level operations vary more between individuals depending on personal background, age, gender, affiliated culture, etc.

1.1 Tactile Haptic Feedback

Obviously, a keyboard that burns or pinches its users' fingers will produce strong emotional reactions. Similarly, more subtle user interactions with *tactile* haptic feedback also combine to strongly influence the user's emotional disposition and attitude. Tactile components include surface texture, shape, and thermal conduction properties of physical computing devices such as media players, pointers, and cell phones. Although these tactile components are frequently considered by usability *designers* in companies, they are often ignored in usability *research*.

1.2 Rendered Haptic Feedback

A less obvious and less explored area of research is the study of user responses to more rendered interactions with physical computing devices. For example, two radio tuning knobs with different levels of friction & inertia may enable a user to tune the same radio station with equal proficiency. However, one knob may produce a much more favorable affective response. We conjecture that exploring users' affective responses to tactile haptic feedback will help guide experimental design and understanding of affective design for rendered haptic feedback.

2 SUMMARY OF AFFECT RESEARCH

Affective and cognitive processes can occur in less than 10 ms, and people are often unaware of the presence of such processes [23]. Furthermore, Zajonc [28] states that affective responses are believed to be *inescapable*, *irrevocable*, *implicate the self*, *difficult to verbalize*, and *often separable from content*.

Many terms exist to classify emotion. Norman [14] uses the terms

- *Visceral:* primary, automatic, unconscious responses (e.g., the computer display is bright, the cell phone ring tone is loud, the stylus is smooth)
- *Behavioral:* also unconscious responses, but are slightly less automatic (e.g., the bright computer display causes surprise, the loud cell phone ring tone is annoying, the smooth stylus is comforting.) We focus on behavioral re-sponses in this paper.
- *Reflective:* responses involving conscious thought and reflection (e.g., I like how clicking this button with the stylus causes the display brightness to increase).

Generally, reflective responses are most influenced by social and cultural attributes, whereas visceral responses have less variability from person to person. For example, a bright computer display will be equally bright for an office worker or a tribal native who has never seen a computer display before. Of course, there is no hard, exact boundary between these levels of emotion.

Visceral responses will vary the least between different people or groups such as office workers, teenagers, or Lithuanians; whereas, reflective responses will vary the most.

Spence [20] suggests that the sense of touch is well suited to perception of differences in emotion. Thus, although performance measures are often dominated by visual and audio feedback, haptic feedback can potentially play a significant role in influencing affective responses. This is an additional motivation for our choice of haptic examples in this paper.

For over 100 years, psychology researchers have consistently reported almost all affect variability to be described by three dimensions [16], [26]. Other researchers have since validated and refined these dimensions. For example, Lang's self-assessment mannequin (SAM) [12] uses the terms:

- Valence (e.g., pleasantness)
- Arousal (e.g., excitement)
- Dominance (e.g., control or prestige)

Self-report measures and biometric recordings are the primary methods of obtaining affective responses. Generally, self-report measures are preferred for analyzing smaller, *relative* differences between stimuli. Biometric measurements are better for *absolute* measurements. For example, with careful, specific instructions, participants can be more easily guided to focus on details of a design (i.e., when making a rating, they can filter out many affects that are of little relevance to the study). Differences between participants' desired and actual interpretations of instructions is one of the major sources of noise in self-reported measures. Although biometric measures are less affected by such misinterpretations, they are more sensitive to the environment (e.g., they are difficult to use in uncontrolled environments such as field studies). Learnt and biological differences will also affect biometric measurement validity.

2.1 Self-Report Measures

Likert-type scales are often used for each dimension. Thus, a participant will typically be exposed to a stimulus for 5-8 seconds, and then be asked to rate valence, arousal, and/or dominance on a scale (e.g., 1-10). Exposure times of 5-8 seconds have been estimated to give participants enough time to experience the stimulus, without giving them time for much conscious thought (i.e., a 'gut' reaction is desired) [12]. Generally, it is believed that approximately half of one's affective judgment variability is along the *valence* dimension, slightly less than half of the variability is along the *arousal* dimension, and most of the small remainder is along the *dominance* dimension. Hundreds of studies, predominantly vision-based Psychology studies have used these scales.

Because *valence* and *arousal* are believed to account for almost all affective variability, Russell et al. [19] proposed and used these as the basis for a two-dimensional *affect grid*, and also related more subtle, specific affective attributes (e.g., *happy*, *sad*, *joy*, *excited*, *frustrated*) to various regions of the affect grid (see Figure 1). Studies measuring more subtle affective states than the main dimensions of *valence*, *arousal*, and *dominance* have had some, but more limited success. For example, attempts have been made to map subtle affective attributes to a defined subregion of a 2D valence & arousal (i.e., affect) grid [11].

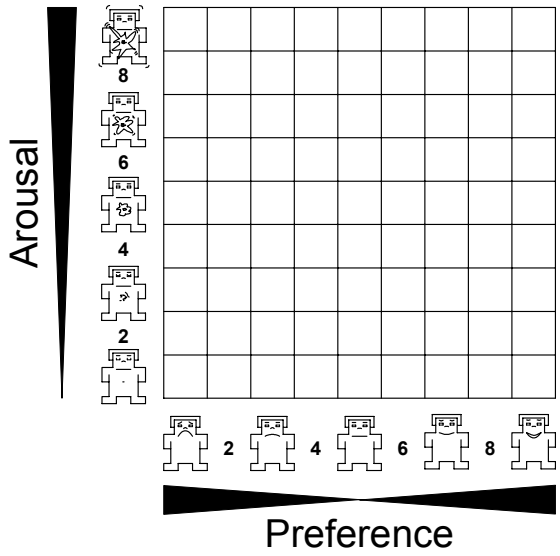

Figure 1: The affect grid. After exposure to a stimulus, participants place an 'X' in a box to self-evaluate their level of valence and arousal (based on Russell et al. [19]).

Several caveats arise from the use of any rating scale. Daniel gives an excellent summary of rating scales for measuring scenic beauty of forest photos [5]. Valence, arousal, and dominance rating scales can be standardized in a similar manner. The two most common problems are:

- Some participants give overly positive or negative responses (i.e., a within-participant average, when compared with others, reveals an individual bias).
- Participants may use the scale differently (i.e., relative differences will vary across participants). To help reduce this effect, experimenters often ask participants to use the full range of the scale for their responses.

2.2 Biometric Measures

Affective responses correlate with a variety of biological responses including changes in muscle tension, skin conduction, heart rate, blood pressure, and breathing rate. Analyses of facial responses have been used by researchers for over 100 years (e.g., Duchenne de Boulogne [6]). More recently, Ekman and Friesen developed the Facial Action Coding System (FACS) where six affective attributes – joy, sadness, disgust, anger, surprise, and fear – can be manually coded from images or video [7]. However, direct measurement with sensors is more accurate and increasingly technically feasible. For example, functional magnetic resonance imaging (fMRI) and use of electroencephalogram (EEG) sensors have been used to monitor brain activity variations for different affective responses [1], [10]. Especially promising new research areas are prefrontal asymmetry and evoked response potentials. Although they accurately record affective responses, fMRI machines are very expensive and their magnetic fields can interfere with many force-feedback interface technologies. Electromyographic (EMG) measurement of facial muscles is often more practical than full-head EEG or fMRI (for reasons of cost, ethics, and complexity). For example, Surakka and Hietanen studied EMG responses to facial expressions [21].

For this paper, we choose an accepted technique for biometric measurement of valence and arousal (e.g., see Conati el al. [4] and Mandryk & Inkpen [13]; Picard also gives a more detailed

discussion of biometric use [18]). *Valence* was measured by calculating the voltage difference between two EMG electrodes placed on the participant's forehead (see Figure 2). *Arousal* was determined from skin conductance (SC) measurements.

3 Affect Evaluation Case Studies

We present two case studies of measuring affective responses to haptic sensations. The first employs a broad range of real tactile stimuli, whereas the second incorporates interaction with haptically rendered virtual environments. Both biometric and self-report measures are explored. Short response times to stimuli are used for both studies in an effort to focus on visceral emotional responses. Such visceral responses vary the least between individuals (i.e., are least affected by gender, age, culture, etc.)

Physical user interfaces typically have a variety of surface properties that can be used deliberately in aid of design. For example, the texture, shape, and heat conduction properties of cell phones, portable media players, and computer mice will influence the user's affective responses and affordances.

The rendered interaction case study shows an example of affective responses to a rendered aspect of a user interface. Specifically, we measure affect in response to different possible knob motions that control access to items on a list (e.g., a radio station frequency or a combo box graphical user interface widget), while the user experience a variety of rendered rendered haptic environments Thus, this case study relates closely to haptic scroll-wheel properties of a mouse used for common computer desktop tasks such as scrolling list boxes or windows.

We hypothesized that the affective differences between the our force-feedback knob renderings would be subtle. Compared to all the emotional stimuli a person experiences in their daily activities, the rendered knobs span a narrow range. Getting hit by a car while walking in a crosswalk, for example, would be much more arousing than the stimuli used in our experiments. Thus, we performed the tactile study as a preliminary test of the efficacy of the self-report and biometric measures.

3.1 Tactile Study

3.1.1 Participants and Apparatus

A total of 9 people (5 male and 4 female) participated. Participants were right-handed and ranged in age from 24 to 33 years ($M = 26.2$, $SD = 2.82$).

Participants sat at a desk while electromyographic (EMG) and skin conductance (SC) traces were logged at 32 Hz using Biograph v. 2.1 software and ProComp+ biometric equipment. Twelve textures believed to span a range of valence and arousal levels were used (see Table 1). Participants were blindfolded while the tactile textures were in front of them, and they were not blindfolded when marking self-reports.

3.1.2 Procedure

The experimenter described the apparatus and procedure to the participants. Two EMG electrodes were then placed on the participant's forehead as shown in Figure 2. Skin conductance sensors were placed on the index and middle finger of the participant's left hand. After each condition, participants were instructed to mark the arousal and valence (i.e., preference) on a scale of 1 to 9 using the affect grid. Participants were asked to try and use the full range of the scales. A sample trial and one complete repetition of all levels were executed to familiarize the participants with the experiment. EMG data was smoothed using a third order low-pass Butterworth filter with 0.1 Hz cutoff frequency. SC data did not need to be smoothed because raw SC data does not contain the high frequency components typically seen in EMG data. A single biometric value was manually calculated by subtracting the baseline voltage from the peak voltage for each stimuli. Accurately and repeatedly determining a baseline voltage is a difficult task that makes biometric measurement inherently uncertain. Arousal and valence scores were calculated by marking the start and end points for each stimulus. Mean voltages were then calculated between these start and end boundaries.

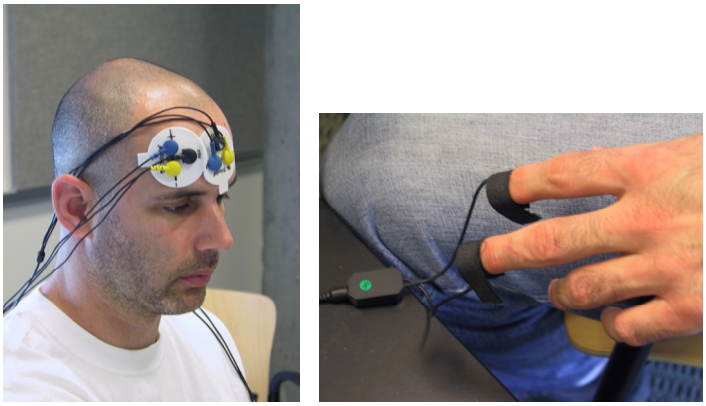

Figure 2: EMG electrode placement on the forehead, and SC electrode placement on the index and middle fingers of the non-dominant (left) hand.

The experimental design used a within-subject factor (*tactile stimulus*) with 12 levels and one repetition.

Table 1: 12 Tactile Stimuli for Tactile Case Study

#	Label	Description
1	FUR	Fox fur
2	GEL	Moist water-based gel
3	PTY	Silly putty™ surface
4	SND	80 grit sandpaper
5	ACR	Acrylic sheet
6	GLS	Glass sheet
7	BSH	Brush with fine plastic tines
8	WD	Maple wooden board
9	OIL	Glass sheet covered in olive oil
10	STK	Double-sided sticky tape on an acrylic sheet
11	HND	Hand stroked by experimenter
12	SLK	Silk

3.1.3 Results

The data were checked for fit to a normal distribution using a Q-Q plot. Normality can be assumed. Figure 3 shows means for the arousal and valence ratings (from the affect grid) for the 12 stimuli listed in Table 1.

For the self-report ratings, significant main effects for the affect grid ratings were found between stimulus and arousal ($F(11, 88) = 10.8$, $p < .001$, $\eta^2 = .574$), and between stimulus and valence ($F(7.14, 57.2) = 10.6$, $p < .001$, $\eta^2 = .571$). A Huynh-Feldt correction for sphericity was used because Mauchly's test for sphericity yielded $\varepsilon = .649$ for valence. No significant main effects were observed for the biometric data; although, several interesting trends were observed (e.g., see Figure 4). Mean EMG and SC voltages for the 12 stimuli are shown in Figure 5.

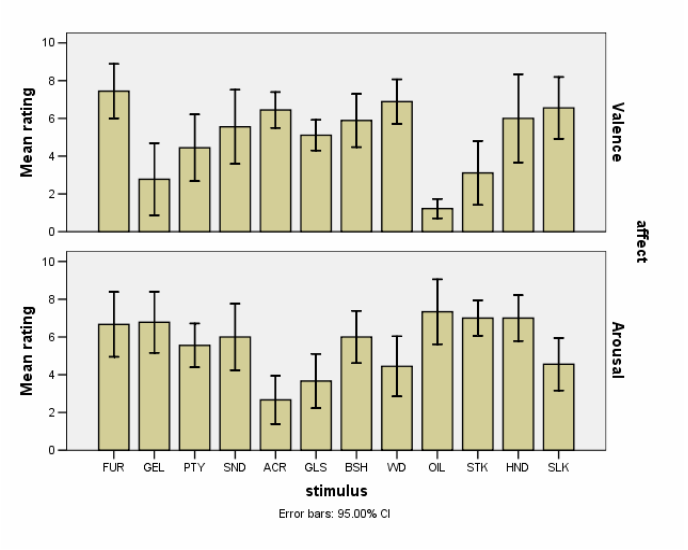

Figure 3: Mean self-reported arousal and valence ratings for 12 tactile surfaces listed in Table 1

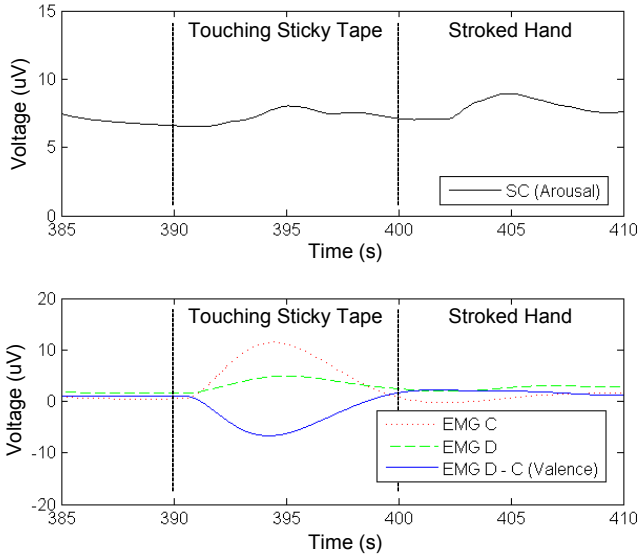

Figure 4: EMG (lower) and SC (upper) data for a participant touching a sheet of double-sided sticky tape, and a participant's hand being touched by the experimenter. High-frequency components of the raw EMG data were smoothed using a third order low-pass Butterworth filter with 0.1 Hz cutoff frequency; the presented SC data are the raw values.

3.1.4 Analysis

Our observation of significant self-reported rating scores, but an absence of significant differences in biometric values was not entirely unexpected. When rating stimuli, participants were asked to use as much of the affect grid as possible. Thus, use of the rating scale yields affect differences in the context of the tactile stimuli observed. Conversely, biometric measurements represent affect differences in the context of all the participants' life experiences and their evolutionary affective predispositions (i.e., affective judgements have both learnt and biological components [8]). Compared to this, the 12 tactile stimuli presented span a relatively small affective range: the difference between feeling

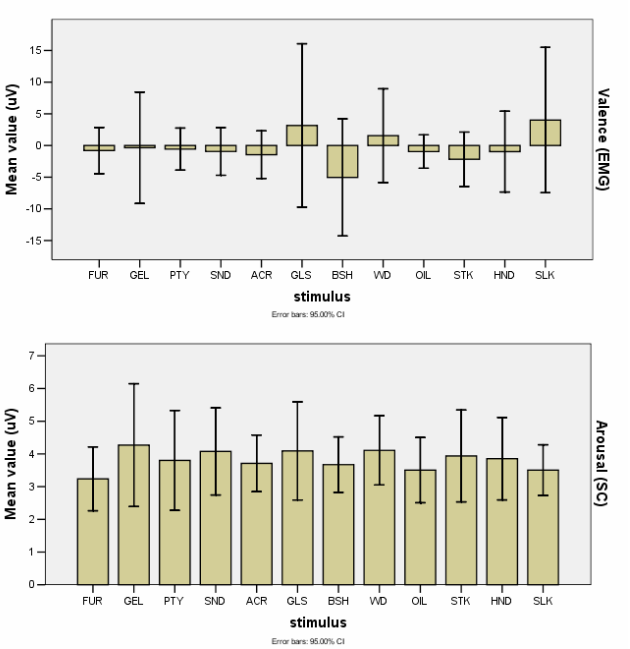

Figure 5: Mean biometric arousal (SC) and valence (EMG) ratings for 12 tactile surfaces listed in Table 1

glass or acrylic pales in comparison to skydiving versus relaxing on the beach. Nevertheless, in the context of a user interface environment, small effects add up. Considering the very high effect sizes of $\eta^2 = .574$ for arousal and $\eta^2 = .571$ for valence for the rating data, one might expect to see significant differences in the biometric data if another study was performed with more participants, or possibly using a different algorithm for calculating arousal and valence scores from the raw biometric signals. (Cohen recommends classifying low, medium, and high effect sizes to be $\eta^2 = .01$, $\eta^2 = .059$, and $\eta^2 = .138$, respectively [3].)

Figure 3 shows arousal and valence rating differences between the 12 different stimuli. Because we observed very high effect sizes for arousal and valence, the results in Figure 3 are strong. The acrylic sheet (stimulus ACR) was rated the least arousing stimulus ($M = 2.7$). Glass (stimulus GLS), wood (stimulus WD), and silk (stimulus SLK) were also ranked as low arousal textures – $M = 3.7$, $M = 4.4$, $M = 4.6$, respectively. Touching a sheet covered in oil (stimulus OIL) received a high arousal rating ($M = 7.3$), and the lowest valence rating ($M = 1.1$). It is interesting to note that there was strong agreement (i.e., low variance) among participants that touching oil was not pleasant (i.e., low valence); participants varied more in their rating of *how* strongly they disliked the stimulus. Ratings of valence for the experimenter touching the participant's hand had a wide variance. Generally, the results in Figure 3 are what we would intuitively expect.

The upper section of Figure 4 shows example biometric measurements of arousal, and the lower section Figure 4 shows example biometric measurements of valence. At about 403 seconds into the trial, an increase in skin conductance (SC) occurs when a participant's hand is touched by the experimenter (starting at 400 seconds into the trial), suggesting an increased level of participant arousal. Skin conductance measurements typically have a 2-3 second lag, and this is exactly what we observe in Figure 4. A less pronounced increase in SC was also observed when the participant touched double-sided sticky tape. The EMG D – C curve is slightly positive for hand stroking case – indicating a slightly positive valence (i.e., preference) for their hand being touched. Conversely, a very strong preference reaction was

observed when the participant touched double-sided sticky tape. The EMG D – C curve dips sharply – indicating a strong negative valence (i.e., dislike) for the sticky tape.

The self-report rating and biometric results also illustrate the point that with rating scales, most subjectivity is from the participant's ability to rate the stimuli. Conversely, with biometric data, most subjectivity is from the experimenter's calculation of arousal and valence scores. For example, there is approximately a 2 second lag between the time a participant becomes aroused, and their skin conductance becomes elevated. Where to start and stop recording voltages values for a particular stimuli, and normalization of stimuli are non-trivial problems.

The error bars of the biometric voltages in Figure 5 are very large. Consequently, we can not draw meaningful relationships between these bar charts and the self-reports in Figure 3. These biometric voltages are given primarily for completeness. They may also suggest the amount of power (e.g., number of subjects) needed for more meaningful future studies involving such biometric measurements.

3.2 Rendered Interaction Study

The purpose of the rendered interaction experiment was to measure affect as a function of a range of rendered haptic environments. Users felt rendered damping, inertia, and detent knob environments with and without the context of a graphical scrolling task. We conjectured that the differences in emotional responses between the haptic renderings used for this interaction study would be subtle and difficult to record compared to the previous tactile textures of the tactile study.

3.2.1 Participants and Apparatus

A total of 15 right-handed people (9 male and 6 female) participated in this experiment; ages ranged from 24 to 27 years ($M = 24.7$, $SD = 1.18$).

Figure 6 illustrates the experimental setup. Participants sat at a desk approximately 50 cm away from a computer display measuring 36 cm wide by 29 cm high, and used their right hand to interact with a force-feedback haptic knob anchored to the desk. Noise canceling headphones were worn to block sounds from the force-feedback dev ice. Visual distractions were reduced by seating the participants at a desk facing a corner of the room.

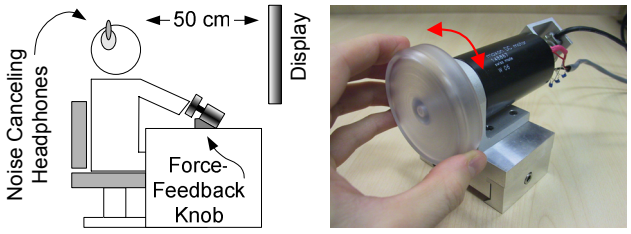

Figure 6: Experimental apparatus

Our custom built force-feedback knob is illustrated in Figure 6. Torques were supplied using a Maxon RE40 DC motor, and position was measured using a MicroE optical encoder operating at 320 000 counts / revolution. A 5000 Hz haptic update loop was coded in C++ using the Real-Time Platform Middleware (RTPM) [17]. OpenGL was used to code the graphical display. The graphical client got the knob position from the haptics server to keep a 60 Hz graphical update rate. We custom built this setup because haptic knob systems capable of rendering such dynamic effects are not yet commercially available.

Three haptic models were used as illustrated in Table 2 with parameter values a_i, b, and m. To improve stability, velocities were low-pass filtered using a tenth order Butterworth filter at a 400 Hz cutoff frequency. Inertia was modeled using a spring & damper virtual coupling to a simulated mass [2].

Table 2: Force-feedback models used

Model	Torque Component	
Detents	$\tau = a_1 \cdot Sin(a_2 \cdot \theta)$	(1)
Viscous Damping	$\tau = b \cdot \dot{\theta}$	(2)
Inertia	$\tau = m \cdot \ddot{\theta}$	(3)

Figure 7 illustrates the graphical display. An almost black background was used (the background had a touch of blue to reduce participant eye strain). A red target value was shown to the left or right of a cyan counter value. Rotating the knob counterclockwise / clockwise would decrement / increment the counter value by 1 unit, respectively. The target value appeared to the left / right of the counter if the target was less / more than the counter, respectively. If the counter equaled the target, the target would appear on both sides of the counter.

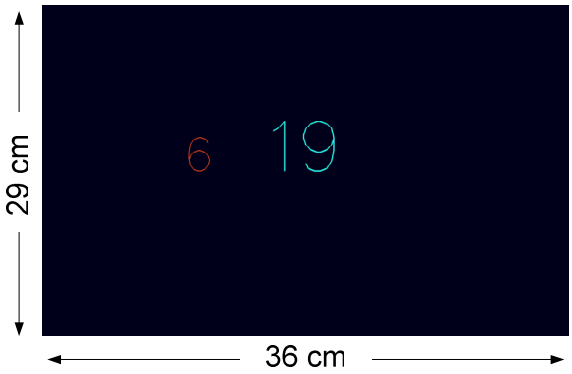

Figure 7: Screen capture of the graphical display

3.2.2 Procedure

The experimental design used two within-subject factors (*feedback* and *knob stimulus*) with two repetitions.

The *feedback* factor had 2 levels:

- *Freeform exploration:* approximately 5 s freeform exploration of different knob models followed by 10 s to record their valence rating followed by a 4 s rest before the next condition.

- *Target finding:* a timed target task taking approximately 5 s (see Figure 7) where the participants rotated the knob until they matched the counter value to the target value. The same 10 s rating time and 4 s rest time allowances as the freeform exploration task were then given. Task completion times were measured to enable performance comparisons.

Because the *target finding* task might have influenced the way participants performed the *freeform exploration* task, all participants performed the *freeform exploration* task first, then the *target finding* task.

The knob stimulus factor had 7 levels of damping, inertia and detents as shown in Table 3 (refer to Equations 1, 2, and 3 for coefficient meanings).

Table 3: 7 Stimuli for Rendered Interaction Case Study

#	Label	b	m	a_1, a_2	Description
1	NON	0	0	0, 0	No force feedback (control)
2	FR1	15	0	0, 0	Small viscous friction
3	FR2	30	0	0, 0	Large viscous friction
4	MS1	0	0.4	0, 0	Small inertia
5	MS2	0	2.5	0, 0	Large inertia
6	DT1	0	0	80, 2π	1 detent / graphic list item
7	DT2	0	0	80, 60	High frequency detents

[Units: b = volts / rad/s; m = volts / rad/s^2; a_i = *volts / rad*]

The experimenter described the apparatus and procedure to the participants. After each condition, participants were instructed to mark the valence (i.e., preference) on a scale of 1 to 9. A sample trial and two complete repetitions of all levels were executed to familiarize the participants with the experiment. Unlike the first study, we only asked participants to rate valence (not arousal) because pilot studies suggested that participants had difficulty assigning different arousal ratings to the 7 different stimuli. Biometrics were measured using the same procedure as study 1.

3.2.3 Results

As in the first study, the data were checked for normality using a Q-Q plot. Normality can be assumed. Figure 8 shows mean self-reported valence ratings of the 7 stimuli listed in Table 3.

We observed significant main effects for task (i.e., freeform exploration or target finding) ($F(1, 14) = 5.75$, $p < .031$, $\eta^2 = .291$) and stimulus ($F(4.48, 46.5) = 5.79$, $p < .001$, $\eta^2 = .293$) as well as a significant interaction between task and stimulus ($F(2.92, 40.8) = 4.89$, $p < .006$, $\eta^2 = .259$). Huynh-Feldt corrections for sphericity were used for the *task* main effect and the *task x stimulus* interaction because Mauchly's test for sphericity yielded $\varepsilon = .746$ and $\varepsilon = .628$, respectively. Furthermore, the final repetition of 7 stimuli was used to calculate all statistics except for reliability statistics where inter-repetition consistency was explored.

No significant main effects were found for the biometric data. Although, promising trends were found as in the first experiment.

Reliability analyses of the results were tested (i.e., how likely we would see the same effects for the valence self-report ratings). Cronbach alpha tests for reliability were performed between the 1st, 2nd, and 3rd repetitions. For the *freeform exploration* task, stimuli FR2, MS1, and MS2 had low Cronbach alpha scores of $\alpha = .61$, $\alpha = .59$, $\alpha = .57$. The other stimuli were all above the recommended value of $\alpha > .70$ [15]. For the *target finding* task, stimulus FR1 had a low Cronbach alpha score of $\alpha = .64$. The other stimuli were all above the recommended value of $\alpha > .70$. These reliability scores (i.e., Cronbach alpha values) suggest an acceptable chance of repeating the observed results for the purposes of this paper. However, replicating the study with more participants (e.g., 100) would be advisable before making knob interaction design decisions.

For practical reasons, favourable affect ratings are often of secondary concern to performance ratings. Consequently, we explore relationships between affect and performance for the *target finding* task. A significant main effect for stimulus was observed ($F(4.78, 66.9) = 5.68$, $p < .001$, $\eta^2 = .288$). Huynh-Feldt corrections for sphericity were used for the stimulus factor because Mauchly's test for sphericity yielded $\varepsilon = .797$.

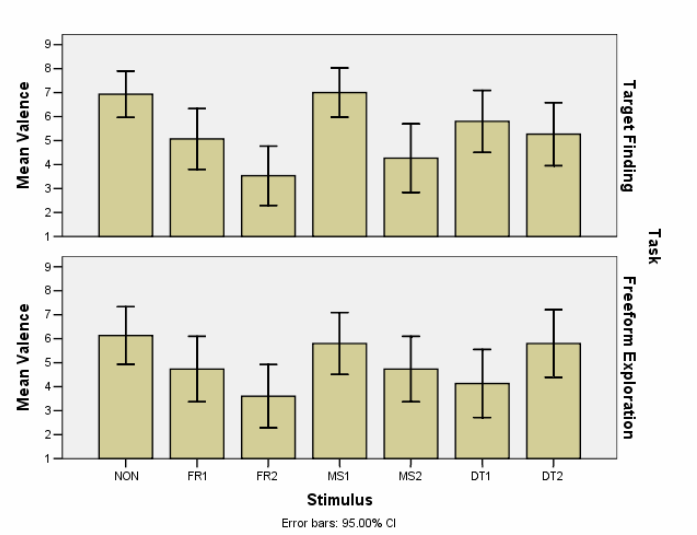

Figure 8: Mean self-reported valence ratings for the freeform exploration and target finding tasks

Significant differences between *freeform exploration* and *target finding* tasks were also found for the NON, $F(1, 14) = 15.68$, $p < .001$, MS1, $F(1, 14) = 5.02$, $p < .042$, and DT1, $F(1, 14) = 14.91$, $p < .002$ conditions.

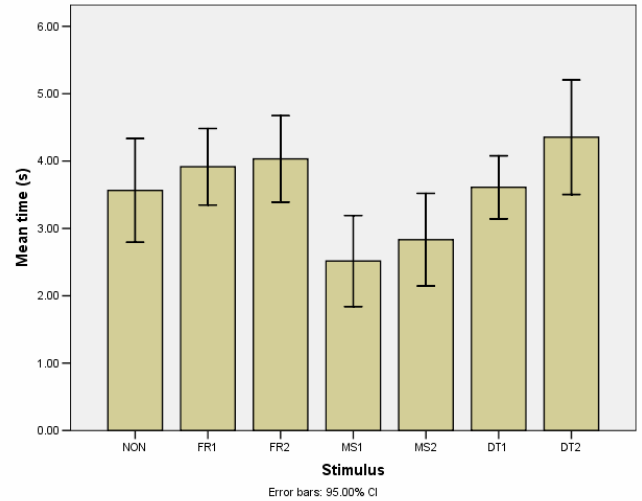

Figure 9: Mean times of target acquisition times for each stimulus in the target finding task

3.2.4 Analysis

As one might expect, giving participants a context in which to evaluate the stimuli occasionally changed their valence ratings. For example, comparing the freeform exploration and target finding tasks, there was a significant *decrease* in valence ratings for the knob with lots of inertia (stimulus MS2), and a significant *increase* in valence for the detent knob (one detent per counter number increase / decrease – (stimulus DT1), and the knob with a small amount of inertia (stimulus MS1).

Looking at the performance data in Figure 9, we can see significant acquisition time differences between knobs with simulated viscous friction (stimulus FR2) and a knobs with

simulated inertia (stimulus MS1). The larger variance for friction stimulus FR2 vs. stimulus FR1 might be explained by variance in physical strength of participants (i.e., a physically strong or weak participant should be able to turn a knob with modest amounts of friction with similar ease). Similarly, during post-experiment discussions, several participants remarked that the larger amount of inertia in stimulus MS2 vs. stimulus MS1 made fine-tuning more difficult at the start and end of the target finding task (e.g., they felt that they were overshooting). Stimulus NON had less friction and inertia than the other stimuli, but there was very large variance between participant times and inconclusive performance differences. This supports the notion that there is a 'sweet-spot' amount of friction and inertia, when positioning performance is at stake.

There is a general trend that knobs that aided a participant's target finding performance corresponded to higher valence ratings. However, a more interesting result is the relatively larger *valence* differences (see Figure 8) between small and large friction levels (stimuli FR1 & FR2), and between small and large inertia levels (stimuli MS1 & MS2), compared to the *performance* differences (see Figure 9). In other words, we observed cases where participants tended to prefer one stimulus over another stimulus *even though* there were minimal performance reasons to make such a preference rating.

4 Conclusions and Future Work

We have demonstrated the use of self-report and biofeedback affective measurement tools for analyzing both tactile and rendered physical user interface components – surface texture and rotary knob movement, respectively. Because the chosen case studies represented relatively subtle differences in affect, self-reported measures tended to produce better results than the biometric measures. Significant differences in valence and arousal levels were observed for different textures. For several rotary knob movements, significantly different ratings of valence were observed for knob movements that helped users perform a simple list scrolling task. Additionally, participants gave knob movements that improved task performance significantly higher valence scores. As one would expect, affect rating reliability improved when participants focused on a particular context. The less intuitive result of finding different affect ratings between different knob movements and different textures *regardless of context* was also shown. (i.e., list scrolling).

Different applications than our two case-studies will undoubtedly require use of different tactile and rendered parameters than those presented in our experiments. Thus, our primary contribution is not the *specific* haptic parameters used in each case-study, but the documentation of typical relative differences for tactile and rendered haptic parameters using accepted self-report and biometric measures.

Future work will include additional self-reported affect studies with a greater variety of haptic interfaces and contexts. For particularly interesting small subsets of interfaces and contexts, biometric studies with more participants and repetitions will be used to explore more absolute affect ratings and individual differences. Other more expressive user study contexts, such as moving a graphical object on a computer screen, could also yield interesting results. Furthermore, more subtle study of weightings between affect versus performance and cost could help motivate more rapid adoption of appropriate affective interfaces into commercial products.

Acknowledgements

The authors would like to acknowledge George Pava for his technical help, and Steve Yohanan for modeling the EMG and SC electrodes in Figure 2. This research was partially funded by the Natural Sciences and Engineering Research Council (NSERC) of Canada and Immersion Corporation.

References

[1] Allen, J. B., Coan, J. A., & Nazarian, M. Issues and assumptions on the road from raw signals to metrics of frontal EEG asymmetry in emotion. *Biological Psychology*, pages 183-218, 2004.

[2] Colgate, J. E. and Schenkel, G. Passivity of a Class of Sampled-Data Systems: Application to Haptic Interfaces, *American Control Conference* (June 29 - July 1, Baltimore, MD), 1994.

[3] Cohen, J. Eta-squared and partial eta-squared in com-munication science. *Human Communication Research, 28*, Oxford Journals, pages 473-490, 1973.

[4] Conati C., Chabbal, R., and Maclaren H. A Study on Using Biometric Sensors for Monitoring User Emotions in Educational Games. In *Workshop on Assessing and Adapting to User Attitudes and Affect: Why, When, and How? User Modeling*, (June 22-26, Johnstown, PA), Springer, 2003.

[5] Daniel, T. C. Measuring the Quality of the Natural Environments. *American Psychologist, 45*, pages 633-637, 1990.

[6] Duchenne de Boulogne, C.-B. *The Mechanism of Human Facial Expression*, Paris: Jules Renard, 1862, (edited and translated by R. Andrew Cuthbertson, Cambridge: Cambridge Univ Press, 1990).

[7] Ekman, P. and Friesen, W. *Manual for the Facial Action Coding System (FACS)*. Palo Alto: Consulting Psychologists Press, 1978.

[8] Ittelson, W.H. Environment perception and contemporary perceptual theory. In W.H. Ittelson (Ed.), *Environment and cognition*. New York: Seminar Press, 1973.

[9] Kahneman, D. and Tversky, A. Prospect theory: An analysis of decision under risk. *Econometrica, 47*, pages 263-291, 1979.

[10] Kemp, A. H., Gray, M.A., Eide, P., Silberstein, R.B., & Nathan, P.J. Steady-State Visually Evoked Potential Topography during Processing of Emotional Valence in Healthy Subjects. *NeuroImage, 17*, pages 1684-1692, 2002.

[11] Killgore, W. D. The Affect Grid: a moderately valid, nonspecific measure of pleasure and arousal. *Psychological Reports, 83*(2), pages 639-642, 1998.

[12] Lang, P. J. The Emotion Probe. *American Psychologist, 50*(5), pages 372-385, 1995.

[13] Mandryk, R. L., and Inkpen, K. M. Physiological Indicators for the Evaluation of Co-located Collaborative Play. In *Conference on Computer Supported Collaborative Work (CSCW)*, ACM, 2004.

[14] Norman, D.A. Emotional Design: Why we love (or hate) everyday things. New York: Basic Books, 2004.

[15] Nunnally, J.C. *Psychometric theory, 2nd Ed.*, 1978, New York: McGraw-Hill.

[16] Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. *The measurement of meaning.* Chicago: University of Illinois Press, 1957.

[17] Pava, G. and MacLean, K. Real Time Platform Middleware for Transparent Prototyping of Haptic Applications, In *HAPTICS*, (March 27-28, Chicago, IL), IEEE, 2004.

[18] Picard,R. *Affective computing*, 1995..Cambridge:MIT Press.

[19] Russell, J. A., Weiss, A., & Mendelsohn, G. A. Affect Grid: A Single-Item Scale of Pleasure and Arousal. *Journal of Personality and Social Psychology, 57*(3), pages 493-502, 1989.

[20] Spence, C. Crossmodal Attention and Multisensory Integration: Implications for Multimodal Interface Design. In *International Conference on Multimodal Interfaces*, (November 5-7, Vancouver, BC), 2003.

[21] Surakka V. and Hietanen J.K. .Facial and emotional reactions to Duchenne and non-Duchenne smiles, *In-ternational Journal of Psychophysiology, 29*(1), pages 22-33, 1998.

[22] Tellegen, A., Watson, D. and Clark, L.A., On the dimensional and hierarchical structure of affect, *Psychological Science, 10*(4), 1999.

[23] Tesser, A. and Martin, L. The psychology of evaluation. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social Psychology: handbook of basic principles*, New York: Guilford Press, pages 400-432, 1996.

[24] Ulrich, R.S. View Through a Window May Influence Recovery from Surgery, *Science*, 224, pages 420-421, 1984.

[25] Ulrich, R.S. Effects of Hospital Responses to Vegetation and Landscapes, *Landscape and Urban Planning, 13*, pages 29-44, 1986.

[26] Wundt, W. *Outlines of psychology*. Leipzig: Wilhelm Englemann, 1907.

[27] Yang, D. and Lee, W. Disambiguating music emotion using software agents. In *International Conference on Music Information Retrieval*, (October 10-14, Barce-lona, Spain), 2004.

[28] Zajonc, R.B. Feeling and Thinking: Preferences Need No Inferences. *American Psychologist, 35*(2), February, pages 151-175, 1980.