Elsevier Editorial System(tm) for Interacting with Computers
Manuscript Draft

Manuscript Number: IwC 2589R2

Title: Affective Responses to System Messages in Human-Computer-Interaction: Effects of Modality and Message Type

Article Type: Regular Paper

Keywords: System Messages; Affect; Physiological Responses; Affective Computing; Usability; Interface Design.

Corresponding Author: Prof. Dr. Hans-Rüdiger Pfister,

Corresponding Author's Institution: Leuphana University Lüneburg

First Author: Hans-Rüdiger Pfister

Order of Authors: Hans-Rüdiger Pfister; Sabine Wollstädter; Christian Peter

Manuscript Region of Origin: Europe

Abstract: Affective responses of users to system messages in Human-Computer Interaction are a key to study user satisfaction. However, little is known about the particular affective patterns elicited by various types of system messages. In this experimental study we examined if and how different system messages, presented in different modalities, influence users' affective responses. Three types of messages, input requests, status notifications, and error messages, were presented either as text or speech, and either alone or in combination with icons or sounds, while users worked on several typical computer tasks. Affective responses following system messages were assessed employing a multi-modal approach, using subjective rating scales as well as physiological measures. Results show that affective responses vary systematically depending on the type of message, and that spoken messages generally elicit more positive affect than written messages. Implications on how to enhance user satisfaction by appropriate message design are discussed.

Prof. Dr. Hans-Rüdiger Pfister, Leuphana Universität Lüneburg
Wilschenbrucher Weg 84, 21335 Lüneburg

To the Editor

Interacting with Computers

Prof. Dr. Hans-Rüdiger Pfister
Institut für Experimentelle
Wirtschaftspsychologie (LüneLab)
Leuphana Universität Lüneburg
Wilschenbrucher Weg 84
21335 Lüneburg
☎ 04131 – 677 7759
pfister@uni-lueneburg.de

-

Dear Editor,

I would like to submit a revised second revision of the manuscript IwC2589 " Affective Responses to System Messages in Human-Computer-Interaction: Effects of Modality and Message Type" to "Interacting with Computers".

A detailed response to the reviewer comments is attached as a separate document.

Sincerely,
Hans-Rüdiger Pfister

***Author Details**

Affective Responses to System Messages in Human-Computer-Interaction: Effects of Modality and Message Type

Hans-Rüdiger Pfister[a]

Sabine Wollstädter[a]

Christian Peter[b]

[a]Leuphana University Lüneburg, Institute of Experimental Industrial Psychology – LueneLab, Germany, Wilschenbrucher Weg 84, D-21335 Lüneburg

[b]Fraunhofer Institute for Computer Graphics, Germany, J. Jungius Str. 11, 18059 Rostock

Corresponding Author:

Hans-Rüdiger Pfister
Leuphana University Lüneburg
Institute of Experimental Industrial Psychology – LueneLab
Wilschenbrucher Weg 84
21335 Lüneburg, Germany

email: pfister@uni-lueneburg.de
phone: +49-(0)4131-677 7759
fax: +49-(0)4131-677 7735

Research Highlights

- Different system messages elicit distinguishable affective responses during human-computer interaction.
- Spoken messages lead to more pleasant affective responses than written text.
- Input requests trigger feelings of pleasantness and dominance, whereas error messages increase arousal and feelings of unpleasantness.
- A multi-modal approach, employing physiological measurements as well as verbal reports, proves to be suitable to assess complex affective responses of users.

Affective Responses to System Messages in Human-Computer-Interaction: Effects of Modality and Message Type

Hans-Rüdiger Pfister[a]

Sabine Wollstädter[a]

Christian Peter[b]

[a]Leuphana University Lüneburg, Institute of Experimental Industrial Psychology – LueneLab, Germany, Wilschenbrucher Weg 84, D-21335 Lüneburg

[b]Fraunhofer Institute for Computer Graphics, Germany, J. Jungius Str. 11, 18059 Rostock

Corresponding Author:

Hans-Rüdiger Pfister
Leuphana University Lüneburg
Institute of Experimental Industrial Psychology – LueneLab
Wilschenbrucher Weg 84
21335 Lüneburg, Germany

email: pfister@uni-lueneburg.de
phone: +49-(0)4131-677 7759
fax: +49-(0)4131-677 7735

Abstract

Affective responses of users to system messages in Human-Computer Interaction are a key to study user satisfaction. However, little is known about the particular affective patterns elicited by various types of system messages. In this experimental study we examined if and how different system messages, presented in different modalities, influence users' affective responses. Three types of messages, input requests, status notifications, and error messages, were presented either as text or speech, and either alone or in combination with icons or sounds, while users worked on several typical computer tasks. Affective responses following system messages were assessed employing a multi-modal approach, using subjective rating scales as well as physiological measures. Results show that affective responses vary systematically depending on the type of message, and that spoken messages generally elicit more positive affect than written messages. Implications on how to enhance user satisfaction by appropriate message design are discussed.

Keywords: System Messages, Affect, Physiological Responses, Affective Computing, Usability, Interface Design.

1. Introduction

For humans it is natural to use a great variety of modalities for communication, such as voice, gesture, and facial expression. However, Human-Computer Interaction (HCI) in common applications such as text processing or spread sheet calculation has largely relied on written textual information when conveying information about the system. In human face-to-face communication, in contrast, speech, prosody, and non-verbal signals are essential to convey information about the speaker's intention, and in particular about mental states such as emotions. Affective signals and responses are of paramount importance in coordinating human discourse. However, to integrate these findings into HCI research and the design of interfaces has turned out to be more complex than expected [1, 2].

We focus on common system messages and assume that typical messages such as status or error notifications not only convey factual information, but might also trigger affective responses. Hence, system messages are double-edged events. On the one hand, they are imperative to ensure efficient interaction and to avoid errors; on the other hand, system messages constitute interruptions which might be experienced as annoying and might be detrimental to performance [3-5]. However, there is little evidence on what type of affective responses are elicited by task-related interrupts, and what kind of messages users actually prefer when interacting with computers.

In this study, we examine basic affective responses to common system messages delivered in different modalities. Affective states not only influence a user's general attitude towards an application, they may also elicit behavioral tendencies on how to respond to messages, for example, by redirecting one's attention, or by increasing one's propensity for risky actions [see 6, 7]. In contrast to other studies, we do not focus on responses to exceptional system states such as failures, or on the issue of how to intentionally modify users' affective states [8], but on the question of how to optimally convey ordinary system

messages, such as error signals or status notifications, in order to increase user satisfaction, and consequently reduce errors and increase productivity.

Affective responses will be considered as a multi-modal construct on two levels, on the level of physiological changes [9, 10] as well as on the level of subjective evaluations [11, 12]. In the following section, we first clarify our approach and terminology. Then, we briefly sketch the various roles of affect in HCI, before describing our research design.

1.1. Affect, Emotion, and Physiological Parameters

We distinguish between affect, emotion proper, and mood [for similar and diverging approaches see 6, 13, 14]. Affect is taken as a general category for any valenced feeling state, be it conscious or unconscious, with or without particular cognitive appraisals, and with or without specific physiological or neuronal correlates. Subjectively, affect is usually experienced as a general feeling of pleasure or displeasure [15-17]. An *affective response* is an immediate and for the most part automatic reaction to an eliciting event, such as a fearful reaction upon seeing a snake, or a fit of frustration after experiencing the third computer crash in a row.

In contrast, we conceive of an emotion proper as a conscious mental state, constituted by particular cognitive appraisals imposing meaning on the currently perceived situation, and directed towards a specific object; emotions such as anger, pride and shame are typical examples [18]. Specific emotions will involve a valenced affective response as a constituent part, similar to what Russell [17] and Barrett [15] call core affect. Mood, on the other hand, is conceptualized as a non-specific and non-directed background feeling of longer duration and lower intensity [14]. Mood can work as a filter or as a disposition to experience particular emotions [19].

A common component of all affective constructs is the notion of valence, that is, any affective response involves an evaluation of the eliciting event with respect to its being pleasant or unpleasant [17]. This basic affective response is assumed to be quick and automatic in most cases [20]. In this study, we will be concerned with this kind of *immediate affective responses of users* caused by standard system messages of various kinds. If users respond affectively to system messages, this may influence further cognitive processes such as attention [7], problem solving, creativity, or motivation [21]. Designers need to take these effects into account when trying to optimize HCI.

In addition to valence, many researchers agree that arousal is a second relevant dimension making up the affective space, constituting the so-called circumplex model [15, 17, 22]. The two dimensions valence (i.e., pleasant versus unpleasant) and arousal (i.e., excited versus calm) are assumed to be largely orthogonal; typical emotions such as anger, fear, or happiness may be located within this circumplex, although a unique mapping of discrete emotions to specific dimensional coordinates has turned out to be elusive. Whereas Russell's model depicts valence as a single bipolar dimension, it should be noted that evidence exists that positive and negative valence might be independent, allowing for the simultaneous experience of pleasant and unpleasant feelings [23].

As an extension of the two-dimensional model of affect, dominance (dominant versus submissive) has been suggested as a third dimension [24]. This tri-partite structure goes back to studies on the semantic differential [25] which consistently found the three factors evaluation (i.e., valence), potency (i.e., dominance), and arousal in people's ratings of various stimuli on the semantic differential. In the context of HCI, the feeling of dominance, implying power and control, might indicate that users consider themselves to be able to achieve their goals and to cope with potential computer problems, possibly indicating a high

degree of self-efficacy [26]. In contrast, low dominance might indicate helplessness, when users feel subjected to the machine as is often the case with novices.

Since the early days of emotion research, many researchers assume a close correspondence between the subjective experience of affective states and the intensity and pattern of physiological reactions [27, 28]. However, to date empirical evidence on the specificity of physiological patterns as signatures of specific emotions is mixed at best [9, 28-30]. Also, the relationship between physiological measures and valence and arousal is somewhat incongruous, in particular in HCI research [31, 32]. Peter and Herbon (2006) present a broad review of studies involving physiological readings and point out that a possible explanation of contradictory findings might be that different stimuli were used in different tasks in most studies, and that the meaning of a physiological response might be highly context specific.

Generally, stronger feelings of arousal are frequently found to be correlated with increased activity of the sympathetic part of the autonomous nervous systems (ANS), involving increased electrodermal activity, independent of valence [11]. Positive valence tends to be correlated with increased heart rate [11] and body temperature [16], but also a decrease in heart rate has been reported [33].

In sum, it is currently not possible to unambiguously map discrete emotions onto distinguishable patterns of physiological parameters, in spite of tremendous efforts in affective computing research [34]. It is, however, possible to gauge general affect via physiological measures such as heart rate or skin conductance, indicating a general valence or activation component of affective responses. It should be noted that physiological measures have been successfully used to assess non-emotional states in HCI, such as stress and mental workload [35], and we do not claim that physiological changes are solely associated with affective responses.

1.2. Affect in HCI and Usability Research

Affect and emotion have for a long time been neglected in HCI research, but during the last decade many researchers have acknowledged that emotion and affect are important factors in human-computer interaction, and an impressive amount of findings has accumulated to date [2, 34, 36-41].

The particular role of affect in HCI can be seen from several perspectives [14, 38]. First, users naturally respond towards computers in an affective way. From this perspective, computers are entities not different from other objects or events with the potential to elicit affective responses under certain circumstances. A computer that does not work, just as a car that does not run as the driver wishes, will cause frustration and anger. We call this the *user perspective*, referring to affective responses particular events exert on the user. More generally, any emotional effects which are caused by computers, for example, aesthetic feelings caused by the delicate design of a laptop or feelings of pride caused by its expensive features, can be subsumed under this perspective [42-44]. Blends of cognitive, motivational, and affective states may also play an important role; for example, Baker et al. [45] demonstrate the prevalence of cognitive-affective states such as boredom and frustration in computer-based learning.

Second, the user might perceive the computer as if it were another human being [41]. We refer to this phenomenon as the *social perspective*, interpreting human-computer interaction as a social situation. Social situations, not surprisingly, are saturated with emotional experiences [1, 46, 47], causing feelings such as anger, shame, and contempt. This applies to the situation of an individual user interacting with a computer in isolation, but might be even more relevant when several individuals communicate or collaborate simultaneously via computers in social networks or virtual groups [48-50].

Third, the computer might even be able to recognize the feelings of the user. The computer might try to monitor the continuous flow of the users' affective states, interpret his feelings, and respond accordingly. This approach originated in research on affective computing [34, 40, 51, 52], aiming at the development of systems that are not only able to assess relevant affective patterns, but also to respond in a way that is perceived as emotional by human users. Following Fairclough [53] we call this the *biocybernetic perspective*. Ideally, the computer would act as if it were a sentient being, though one of the more benevolent and caring types. In a full-fledged biocybernetic cycle, the computer needs to understand the user's emotion, and it must be able to respond in an adequate manner. Both aims pose severe problems, and despite enthusiastic attempts neither the problem of automated emotion recognition [32, 54-58] nor the problem of system initiated emotion feedback and regulation [8, 59] has at present been effectively solved [51, 60].

In this study, we focus on the user perspective, that is on normal events such as system messages in HCI and examine the affective responses these events elicit in humans. Findings about the relationship of system messages and affective responses might provide information about conditions fostering positive responses. This, in turn, might provide means to increase user satisfaction by delivering messages in an appropriate mode, and thus contribute to the design of adaptive feedback procedures as part of the biocybernetic cycle [61].

1.3 Research Questions

We assume that computer users strive for positive affect or at least try to maintain a dispassionate state, that is, they prefer states of dominance (system control), of low or moderate arousal (calmness), and of positive valence (pleasantness) with respect to their current interaction with the system; correspondingly, they try to avoid states of low

dominance (helplessness), high arousal (nervousness), and negative valence (unpleasantness). This assumption might not apply to the use of computers for entertainment, such as computer games [62, 63], when high arousal, excitement, and, generally, the experience of intense affect is a primary objective. Here, we focus on ordinary office applications, and assume that most system generated stimuli will be of rather low affective intensity, and the user's main objective is to successfully perform his or her task without being interrupted by annoying messages.

When using a typical computer application, such as a word processing or spreadsheet program, a user's affective state will be influenced by a variety of external and internal stimuli. Among others, the ongoing process of achieving one's task, for example, writing a report, will be a major determinant of affect [5]; furthermore, incidental events such as a ringing phone, or an interruption by one's colleague will elicit affective responses. While cognitive effects of interruptions on task performance have been studied since many years [4, 64], their influence on affective processes in the user has been neglected. In this study, we investigate breaks in the form of messages conveyed by the computer while the user tries to accomplish her or his task. Incidental interruptions usually hinder one's progress, leading to unpleasant feelings [3-5]. However, there is still a lack of evidence on what type of affective responses are caused by system interrupts, in particular, by common system messages.

This study focuses on three related research questions:

Question 1. Do different types of system messages cause specific affective responses?

Question 2. Is the affective impact of system messages moderated by the way or mode they are delivered?

Question 3. Are there differences between affective responses at the level of subjective experience and at the level of physiological changes?

Concerning Question 1, three types of system messages will be studied: *input* messages, *status* notifications, and *error* messages; details will be given in the method section. We consider these messages to be common and normal events in everyday computer applications; we are not interested in rare events such as fatal errors or safety-critical incidents. We will refer to this aspect as *message category*.

Concerning Question 2, it is essential to know the conditions that elicit and moderate affective responses. A plausible assumption is that the mode of communication, that is, written (visual) or spoken (acoustic), is a key moderator; we will refer to this aspect as *global modality* in this paper. In particular, we assume that voice in contrast to text will lead to more pleasant responses. For example, Qiu and Benbasat [65] examined the effect of text-to-speech voice in e-commerce systems and found that text-to-speech fosters consumers' feelings of flow, that is, a pleasant and playful feeling concerning one's interaction with the system [see also 66, 67, 68].

Moreover, in addition to the basic textual message, further information may be presented in combination with the message content, such as a sound or an image. An annoying sound might lead to more arousal and negative affect, whereas an additional image might be less surprising and disturbing, triggering less unpleasant responses. Providing a richer set of symbols has been assumed to increase vividness and social presence, and thereby generate a more positive user experience; however, empirical evidence supporting this assumption is still preliminary [69, 70]. In this study three variants will be examined: Presenting the system message *alone*, presenting the message in combination with an *icon*, and presenting the message in combination with a *sound*. We will refer to this aspect as *specific modality*, and assume that providing a richer set of symbols will foster positive affect.

Concerning question 3, the relationship between verbal reports about the subjective affective experience and physiological measures is still unclear. Hence, subjective ratings as well as a set of physiological parameters will be assessed in order to explore their connections in the context of HCI in standard applications.

2. Method

In an experiment we varied message category, global modality, and specific modality. Three kinds of system messages were used to elicit affect: input requests, status notifications, and error messages. Messages were presented either visually, i.e., as text, or acoustically, i.e., as speech, as well as presented either alone or in combination with an icon or a sound. In order to control for task effects, participants were required to work on three typical computerized tasks, a spreadsheet calculation, an information search, and a typing training task. Affective responses were measured using physiological indicators as well as subjective rating scales.

2.1. Participants

Fifty-four students (41 female, 13 male, age from 19 to 38 years with a mean of 24.9 years) participated in the study. Participants obtained credits as part of their study requirements.

2.2. Design

A three-factorial design with global modality, specific modality, and message category as experimental factors was applied. The factor global modality was varied on two levels, either as a written (i.e., visual modality) or a spoken (i.e., acoustic modality) message. The factor specific modality was varied on three levels, that is, the message was either presented alone, or the message included an appropriate symbol (an icon) in addition to the written or spoken message, or the presentation included an additional alerting sound (a beep).

The factor message category was varied on three levels: The message was either an input request, that is, the user was required to enter information, or a status notification, informing the user about some state parameter of the system, or an error message, informing about an erroneous input of the user (Figure 1).

Global modality was realized as a between-subjects factor, whereas specific modality and message category were varied within-subjects, yielding a 2 (global modality) × 3 (specific modality) × 3 (message category) factorial design with repeated measurement on specific modality and message category. Thus, each participant received all $3 \times 3 = 9$ combinations of message category with specific modality.

Figure 1

In order to prevent artifacts due to task demands and to augment generalizability, participants worked on three different tasks as described abov: A spreadsheet calculation, an information search task, and a typing training task. During each task, all three kinds of message categories were presented at predefined processing steps (Figure 2).

Figure 2

In order to balance task effects, the particular message conditions were assigned to different tasks for each participant. Three experimental sets with different mappings of message-condition to task were constructed, balancing all within-subject factorial combinations across tasks. For example, a participant assigned to experimental set one received the simple written or spoken input request during the typing training, the input plus symbol request during the information search task, and the input plus sound request was

presented during the spreadsheet calculation task. Likewise, if the error message is presented alone when performing the typing task, input request and status notification messages were combined with a symbol or with a sound, respectively. In the subsequent trial, the error message was then combined with a symbol or sound, and input and status message were accordingly presented alone or in another combination (see Figure 1).

Furthermore, order of tasks was balanced across participants, that is, participants either worked on the typing task first, followed by the calculation task, and finally on the search task; or they began with the calculation task, then the search task, followed by the typing task; or beginning with the search task, followed by the typing and calculation tasks.

In sum, order of tasks, order of message categories, and order of specific modality conditions was systematically balanced across participants, thus avoiding position effects and task artifacts. Global modality, however, was varied between two independent groups in order to avoid interference between spoken and written messages (for example, waiting for a voice when only a text is presented).

2.3. Measures

Subjective ratings of affect. Subjective affect assessments were obtained using an assessment screen which displayed the Self Assessment Manikin (SAM) scale introduced by Bradley and Lang (1994), a reliable and valid method to measure affect via self-report [12, 71]. The SAM is a non-verbal pictorial assessment technique that directly measures three affective dimensions: valence, arousal, and dominance. Participants indicated their current affective state online on the screen by a mouse click on a scale mark located below and between the figures, yielding a nine-point rating scale (1 = unhappy, low arousal and dominance; 9 = happy, high arousal and dominance).

The assessment screen was presented three seconds after a target event (e.g., an error message) happened, and disappeared as soon as the participant had rated his or her feelings

on each scale. The three self-report scales were presented once for every system message presentation, so that the working process was interrupted three times. In the typing task, however, because of frequent unplanned typing errors triggering an error message, the assessment procedure was applied only every seventh message.

Physiological measures. The physiological parameters assessed were skin conductance, heart rate, and skin temperature. These measures were chosen because of their prominence in affect and HCI research [35, 72], and, most important, because they are relatively unobtrusive. Many approaches involving physiological sensing to infer affective states have been proposed to collect physiological data, but all imply a trade-off between accuracy and completeness and leaving room for natural unobstructed interaction [32, 73-75]. Unfortunately, usability and comforts of these devices is often deficient. Particularly when investigating mild affective responses, the equipment chosen should be as unobtrusive for participants as possible. These considerations lead us to focus on skin conductance and skin temperature, which are also the preferred parameters in many HCI-related studies, and heart rate as a further, slightly more obtrusive measure. Due to their obtrusiveness, measures such as EEG or fMRI are considered less useful when HCI-related applications are of interest. Devices for heart rate and skin temperature measurement were available as an integrated measurement system from the Fraunhofer IGD research institute, Rostock, Germany (EREC; [73]; electrodermal activity was measures using the Varioport system from Becker Meditec™. Physiological recordings were taken continuously during the complete duration of working on each task.

Skin conductance (SC), indicating electrodermal activity (EDA), is usually considered as a reliable indicator of an affective response [76, 77]. Skin conductance has been found to correlate with arousal, elicited by stimuli that are either positively or negatively valenced [11, 78]. Following recommendations of Dawson et al. (2007), phasic skin conductance was

measured within a three-second interval after stimulus-onset. The skin conductance response (SCR), serving as dependent variable, was computed as the difference between the post-stimulus maximum amplitude and the level of skin conductance at stimulus onset. Skin conductivity was measured using the Varioport Measurement System from Becker Meditec™ by placing two Ag-AgCl sensors on the right food plantar in order to avoid interference with finger activity. The SCL coupler of the Varioport System applied a constant 0.5 Volts across the electrodes, changes were measured in units of μS (microSiemens).

Heart rate (HR) is another indicator for physiological activity elicited by affective stimuli, and one of the most common, reliable and valid method for assessing cardiovascular changes [79]. Heart rate has been shown to be correlated with pleasantness of affective stimuli [11, 13], with pleasant stimuli being associated with increased heart rate. Heart rate was measured using sensors secured to the rib cage with the non-commercial EREC system from the Fraunhofer IGD research institute, Rostock, Germany (for details see [73]). Phasic change in heart rate (measured in bpm) served as dependent variable. It was computed as the difference between the maximum post-stimulus value and a pre-stimulus value. The pre-stimulus value was computed as the average during a five second period previous to stimulus onset. The post-stimulus value was determined as the maximum response within a five second interval after stimulus onset.

Skin temperature (ST) changes constitute a somewhat delayed and less reliable indicator of physiological activity [76]. Venables and Christie [80] showed that decreased temperature may lead to extended time parameters, e.g. recovery time, of SCR. We measured finger temperature using a temperature sensor fixed on the middle finger of the nondominant hand; measurement equipment and software was again provided by the Fraunhofer IGD research institute, Rostock, Germany [73]. Maximum change in skin temperature (in °C)

within a five second interval after stimulus-onset, relative to the average computed across five seconds prior to stimulus-onset, served as dependent variable.

2.4. Procedure and Material

The experiment was conducted individually with each participant. After participants arrived at the laboratory, they were randomly assigned to one of the global modality conditions, as well as to one of the three experimental sets (order of task/message presentation). Participants were then seated in front of the computer, and sensors and chest strap were attached and connected. When the technical functioning of the physiological measurements had been verified, participants were informed about the following tasks and were asked to complete questionnaires about computer experience and demographic variables. All instructions during the experiment were automatically presented at the appropriate stages via the computer interface. During the tasks, system messages were presented following the schema described above (Figure 2). Following each single system message, participants answered the three SAM-assessments, which appeared in the middle of the screen three seconds after the system message, and then continued to work on the task again.

The three tasks were (a) to perform a typing training on keyboard use with a simple word processing program, (b) to correct calculation errors in a spreadsheet program representing fictitious employee travel costs, and (c) to search for addresses of fictitious clients in a data base program; order of tasks was balanced across participants. Each task required approximately five minutes to complete

During each task, participants received three system messages. Input requests required an immediate activity from the participant, fulfilling the request. Status notifications informed about the status of the current task, but no immediate activity was required. Error

messages implied indirect activities, because participants might be asked to correct a previous erroneous activity. For example, during keyboard training, an input request would ask to enter one's user name, a status notification would inform the participant that the practice lesson is finished, and an error message would inform that he or she made a typing error. During the calculation task, the input request would ask if the participant wants to save changes, the status notification would inform about implied changes in travel costs, and the error message would signal that changes need to be made via a specific formula tool. Depending on condition, each message was either presented alone, or in combination with an icon, or in combination with a beep sound. As described above, order of message category and order of combination with icon or sound was balanced across trials (Figure 1).

In the condition with spoken messages, the message was presented by a female voice. Pitch was artificially reduced, yielding a fairly neutral voice. Note that voice per se can be an affective signal [81], but this was not controlled in this study.

3. Results

As outlined above, six dependent variables were analyzed as indicators of affect, three of which are physiological measurements (heart rate, electrodermal activity, and skin temperature), and three of which are subjective ratings of affective feelings (valence, arousal, and dominance). We report analyses separately for each dependent variable with respect to the independent variables global modality (written or spoken message presentation), specific modality (simple message, message plus symbol, or message plus sound), and message category (input request, status notification, or error message).

All analyses are based on a sample of 54 participants, 27 participants for each condition of global modality; with repeated measurement on specific modality and messages category with three levels each, yielding 486 observations per dependent variable. A total of

around four percent missing observations occurred due to technical failures or non-responding participants. Since these missing values were randomly distributed across conditions, they were replaced by the median of the total sample in order not to loose statistical power due to listwise deletion, and in order to keep a balanced design, which is a recommended trade-off if less than five percent of data are missing at random [82].

Outliers of physiological variables were defined as values more extreme than two standard deviations and discarded. Note that the physiological variables are not normally distributed, and sphericity is violated for most within-subjects effects. Thus, for all analyses it was checked if results change after a Greenhouse-Geisser correction, and for all significant main effects a non-parametric test (Kruskal or Friedman test) was performed. Since statistical conclusions did not change according to these checks, we only report traditional ANOVA results.

Concerning the subjective affect ratings, the normality and homogeneity of variance requirements of ANOVA are met; if violations of sphericity occurred it was checked if conclusions change after Greenhouse-Geisser correction which was not the case. It has also been argued that subjective rating scales should be treated as ordinal scales; hence, we conducted parallel tests using a proportional odds ordinal regression model treating the response variable as an ordered categorical scale [83]. Results from these ordinal analyses were virtually identical to the ANOVA analyses, hence we only report the traditional ANOVA tests.

For an overview, all means are shown in Table 1, and all ANOVA tests are summarized in Table 2.

Table 1

Table 2

3.1 Physiological Measurements

Heart rate change. Change in heart rate as response to the message event was analyzed by a 2 (global modality) × 3 (specific modality) × 3 (message category) analysis of variance with repeated measurements for specific modality and message category, and controlling for experimental set. A significant main effect was found for message category ($F(2, 104) = 5.57$, $p = .005$, η^2_p (partial eta-squared) = .11). Heart rate increased significantly for input requests in contrast to status ($p = .008$) and error messages ($p = .009$) according to post-hoc multiple comparisons with Sidak correction; status and error messages did not differ significantly. Although there was a noticeable tendency for greater heart rate changes when messages were spoken in contrast to written messages, this difference was not significant (Figure 3a). Only the simple effect of global modality for the input request condition showed a significant increase in heart rate for spoken in contrast to written messages ($F(1, 50) = 4.66$, $p = .036$). Specific modality showed no effect on heart rate change. In sum, heart rate increases when an input request occurs, and this increase is somewhat stronger for spoken messages.

Figure 3

Skin conductance response. Because of a large proportion of SCR values being zero or close to zero, indicating a non-response, we decided to dichotomize this non-normally distributed variable. All phasic responses smaller than 0.05 μS were coded as non-responses,

and all values larger than 0.05 μS as responses elicited by the stimulus event [77]. A logistic regression model with this dichotomized response as dependent variable, the experimental conditions as predictors, and participants as a random factor to control for repeated measurement, yields a significant effect of message category, $\chi^2(2) = 23.26$, $p < .001$, with the error category yielding significantly larger skin conductance responses ($z = 3.26$, $p = .001$) in contrast to input or status messages (Figure 3b). Global modality yields a marginal effect, $\chi^2(1) = 3.59$, $p < .058$, indicating that with spoken messages SCR decreases slightly. No effect for specific modality was found. These findings are confirmed by a 2 (global modality) × 3 (specific modality) × 3 (message category) analysis of variance with repeated measurements for specific modality and message category, and controlling for experimental set, yielding a significant effect of message category ($F(2, 104) = 9.42$, $p < .001$, $\eta^2_p = 0.18$). Increase in skin conductivity was significantly greater after error messages in contrast to status notifications ($p < .001$) and to input requests ($p = .003$), input and status messages were not significantly different (all post-hoc multiple comparisons with Sidak adjustment of p-values). No effect for specific modality nor for global modality was found. It has also been proposed to normalize the phasic SCR with respect to the tonic base level [77]; an analysis of variance with the normalized values as dependent variable again yields a significant message category main effect, $F(2,98) = 5.91$, $p = .004$, $\eta^2_p = 0.12$.

Skin temperature change. Skin temperature was analyzed by a 2 (global modality) × 3 (specific modality) × 3 (message category) analysis of variance with repeated measurements for specific modality and message category, and controlling for experimental set. The effect of message category turned out to be marginally significant ($p = 0.10$), with input requests leading to a slightly higher skin temperature. Also, a marginal interaction effect between global modality and message category ($p = .11$) as well as between specific modality and message category ($p = .12$) could be detected. Only the simple effect of global modality for

the input condition turned out to be significant ($F(1, 50) = 4.18$, $p = .046$; see Figure 3c), showing a higher average in skin temperature when input requests were spoken in contrast to being written. Altogether, however, skin temperature appears not to be particularly responsive with respect to system messages.

A multivariate analysis confirms the previous findings. With heart rate change, electrodermal activity, and skin temperature combined as a multivariate vector of three dependent variables, and applying a multivariate mixed model analysis [84], we find a significant effect of message category ($\chi^2(1) = 7.95$, $p < .019$, according to a likelihood-ratio test comparing a Null-model containing only experimental set as a covariate, with a model containing message category as an additional predictor); no further main or interaction effects turned out to be significant.

Summary of physiological parameter changes. Phasic heart rate and skin temperature changes show a similar response pattern to messages events, with a clear tendency to increase after input messages in contrast to status notifications and error messages. Electrodermal acitivity shows a reverse pattern, with significantly increased skin conductance following error messages. Furthermore, there is a particular tendency for spoken messages, in contrast to written messages, to increase heart rate and skin temperature following input requests. No effects could be detected with respect to specific modality. In sum, physiological responses are clearly sensitive to particular message categories: Heart rate and skin temperature increase, maybe indicating a neutral or slightly positively valenced orienting response, whereas increased skin conductance presumably indicates heightened arousal after error messages [11].

3.2. Subjective Affective Ratings

Valence. A 2 (global modality) × 3 (specific modality) × 3 (message category) analysis of variance with repeated measurements for specific modality and message category was conducted with subjective valence ratings as dependent variable, controlling for experimental set. The main effect of global modality turned out to be statistically significant, $F(1, 50) = 5.15$, $p = .028$, $\eta^2_p = .10$. Spoken messages obtained higher mean valence ratings than written messages ($M_{spoken} = 5.43$, $M_{written} = 4.94$); in particular, the simple effect between written and spoken messages for the status message condition was significant, $t(52) = 2.56$, $p = .007$. Also, the main effect of message category on valence ratings was significant, $F(2, 104) = 11.76$, $p < .001$, $\eta^2_p = .23$, with input messages obtaining higher ratings than status or error messages (post-hoc comparisons with Sidak correction showed significant differences between input and status messages, $p = .002$, and between input and error messages, $p < .001$; see Figure 4a).

Figure 4

Arousal. A 2 (global modality) × 3 (specific modality) × 3 (message category) analysis of variance with repeated measurements for specific modality and message category, controlling for experimental set, and subjective arousal rating as dependent variable revealed a significant main effect for message category ($F(2,104) = 5.75$, $p = .004$, $\eta^2_p = .11$). According to post-hoc comparisons (with Sidak correction for multiple comparisons), subjective arousal is rated as significantly higher following error messages compared to input ($p = .017$) and status messages ($p = .025$) (see Figure 4b). Though there appears to be a slight tendency for written messages to elicit more arousal, this difference is not significant.

Dominance. A 2 (global modality) × 3 (specific modality) × 3 (message category) analysis of variance with repeated measurements for specific modality and message category, controlling for experimental set, and subjective dominance rating as dependent variable revealed a highly significant main effect for message category ($F(2,104) = 17.09$, $p < .001$, $\eta^2_p = .33$). The message category effect is clear cut: Feelings of dominance are significantly stronger following input requests ($p = .005$) in contrast to status notifications, which are in turn significantly larger compared to error messages ($p = .011$; means are 5.07, 4.62, and 4.30 for input, status, and error messages, respectively; see Figure 4c). The effect of global modality, though showing a noticeable tendency of spoken messages to trigger stronger feelings of dominance, does not reach significance. No effect of specific modality was detected. The pattern of dominance ratings is similar to the valence rating pattern, though the global modality effect is not significant.

A multivariate analysis confirms the previous findings. With valence, arousal, and dominance combined as a multivariate dependent vector, and applying a multivariate mixed model analysis [84], we find a significant effect of global modality ($\chi^2(1) = 5.24$, $p = .022$, according to a likelihood-ratio test comparing a Null-model, containing experimental set as only predictor, with a model containing global modality as an additional predictor), as well as a significant effect of message category (($\chi^2(1) = 21.57$, $p < .001$); no further main or interaction effects turned out to be significant.

Summary of subjective affective ratings. The analysis of subjective ratings of affective responses, measured three seconds after a system message event, yields a typical pattern (Figure 4). Spoken messages tend to elicit larger positive ratings, that is, more pleasant and more dominant feelings, than do written messages. Also, for valence as well as for arousal and dominance, the message category has a strong impact: Input requests are, all in all, evaluated as more pleasant than status and error messages; an analogous pattern applies to

dominance, respectively. Error messages, not surprisingly, yield the least pleasant and least dominant feelings, but elicit stronger arousal.

For a proper interpretation, the base line measurements of affective ratings before starting the experimental trials must be taken into account. Formally, the neutral midpoint of the nine-point bipolar rating scales is five. The empirical base line ratings of valence (M = 6.41, SD = 1.35) as well as of dominance (M = 5.37, SD = 1.39), measured at the start of the experiment before responding to any target events, are well above the average of ratings elicited as response to a message event during experimental trials. That is, positivity of affect generally drops once participants start working on the experimental tasks. Presumably, this is due to the experimental setting and demand characteristics, which for most participants do not constitute a pleasant overall experience.

The arousal ratings shows a somewhat individual pattern, partly opposite to the valence and dominance ratings. While subjective arousal is generally more or less in the vicinity of normal base line arousal (M = 4.56, SD = 1.63), it increases when error messages are presented. Whereas valence and dominance show a common response with respect to message category and global modality, arousal shows an opposite response with maximum intensity following error messages. Arousal is also not affected by global modality. Apparently, subjective arousal as assessed in this study might indicate a variant of negative affective activation, caused merely by annoying error messages.

3.3 Relationships between Subjective Ratings and Physiological Measures

In order to examine the relationships among the measures of affective responses, in particular among physiological and subjective variables, Table 3 shows the correlations among all six dependent variables. Whereas all subjective ratings are highly and significantly correlated, showing a strong positive correlation between valence and dominance, and a

strong negative correlation of both valence and dominance with arousal, the physiological measures are virtually uncorrelated among each other. In particular, no correlations are found between physiological parameters and ratings.

Table 3

This is in contrast with established findings that skin conductance should be at least somewhat positively related with arousal, and heart rate with valence [11]. Arousal and skin conductance do show a small positive correlation in the expected direction (r = .19), but it is not significant (p = .172). Also, valence and skin conductance are somewhat negatively related (r = -.18, p = .198), suggesting that in this particular study it is mainly negative stimuli that trigger electrodermal activity; the large negative correlation between valence and arousal ratings (r = -.59) confirms this conjecture. Also, interestingly, dominance is strongly and positively related with valence (r = .67), indicating that system messages such as input requests and status notification might increase feelings of control and efficacy of users.

In sum, physiological indicators of affect show a systematic pattern as a function of message category and global modality, but their relationship with subjective reports of affect is less clear. We can think of two possible reasons. First, ordinary system messages constitute very weak affective stimuli; as a consequence, automatic visceral responses will be very weak also. As discussed previously when analyzing the SCR measures, we must assume that in many instances the message stimulus did not trigger an affective response at all. Second, it is still unclear how general findings about relationships between physiological parameters and subjective ratings actually are. What has been found when using a large variety of affective pictures or sounds, may not generalize to highly specific contexts such as users

interacting with standard computer applications. Both reasons might contribute why the correlations among physiological and rating variables are virtually zero in this experiment.

4. Discussion

4.1 Summary of Findings

To summarize, we found that the assumption that users experience considerably different affective patterns in response to system messages could be largely confirmed, yielding a positive answer to our research question one. Different message categories such as input requests, status notifications, and error messages elicit different patterns of affective responses. This is found with respect to subjective ratings as well as to physiological parameters. Input requests increase heart rate and skin temperature, and error messages increase skin conductivity. Input requests and to a lesser extent status notifications lead to more pleasant and more dominant feelings compared to error messages, which are associated with unpleasantness and feelings of relatively low dominance. Error messages, on the other hand, increase feelings of subjective arousal. Note that these findings are relative comparisons contrasting different message categories; compared to the pre-experimental base line, ratings of valence and dominance generally drop during experimental trials. A plausible explanation is that performing the experimental trials, in particular being continuously assessed by a variety of instruments, is slightly stressful and unpleasant, and the particular tasks are probably boring for most participants. Additionally, being interrupted constantly during task performance by system messages might be an annoying experience per se.

It turns out that subjective ratings and physiological changes are not correlated in this experiment, giving a preliminary answer to research question three. As can be seen in Table 3, the correlations between heart rate, temperature, and skin conductance on the one hand, and the ratings of valence, dominance, and arousal on the other hand, are virtually zero. Not

surprisingly, valence and dominance are correlated positively, whereas arousal is negatively correlated with valence and dominance. This interpretation conforms with our assumption that dominance is a positively valenced feeling of control and self-efficacy, whereas arousal signifies a negatively valenced feeling of irritation. On the other hand, the fact that skin conductance and heart rate do not correlate as expected from previous research, might be attributed to the particular context, that is, very weak stimuli in an ordinary computer task.

Are users' affective responses sensitive to the particular way system messages are presented? Our findings referring to research question two suggest that the modality used to convey the information, that is, as a written text or as a spoken communication, does indeed to a certain extent influence affective responses. By and large, spoken messages trigger more pleasant affect and are associated with feelings of control, whereas written messages are, in comparison, less pleasant, associated with less control and with higher physiological agitation. However, this finding needs to be qualified as the effect of what we referred to as global modality depends on the type of system message. Input requests and, to a lesser degree, status notifications, are susceptible to changes in modality, whereas error message appear to be immune to this kind of modality change.

What we referred to as specific modality, that is, a combination of messages with symbols or sounds, did not yield any significant effect on affective responses. It remains to be seen if variations in iconic representations, for example, using enlarged icons with explicit affective content, or variations in sound might be more effective. However, we conjecture that variations in iconic or acoustic content or intensity would most likely lead to higher arousal, more irritation, and generally to more unpleasant responses, as users rarely like to be interrupted and conceive of system messages usually as an unpleasant disruption. How to minimize annoyances stemming from technically mediated interruptions might be a research issue in its own right [4, 85].

4.2 Limitations of the Study

Though we did our best to ensure internal validity by balancing treatments across different tasks and participants, a number of limitations remain to be addressed. The sample constitutes a convenience sample of psychology students, so generalizing to other populations should be made with care. More important, the selection of experimental tasks and the selection of message categories represent only a small sample of possible tasks and messages. It may well turn out that affective responses change systematically with tasks and message categories not examined here. In particular, the operationalization of a specific message implies several design decisions; for example, length of message, size and font of text, and kind of voice for auditory messages. It is well known that people react strongly and affectively to human voices [67], and variation in the quality of the voice conveying the message is a good candidate to manipulate affect.

It should be noted that for some participants skin conductance and skin temperature yielded fairly flat measurement profiles, indicating either low or delayed responsiveness to target events. As this reduces systematic variance, statistical results will be rather conservative, and one might presume that true physiological activity is stronger than what could be measured. A phasic analysis might be limited when the affective stimuli are very weak, as is to be expected in office software, in contrast to applications such as computer or adventure games [86].

With respect to the instruments used to measure affective responses, the usual caveats apply. Subjective ratings of affect should be interpreted with caution, since peoples introspective abilities are questionable, and the semantics of a pictorial scale such as SAM might be heterogeneous among participants; also, tendencies such as self-serving bias and social conformity might reduce validity. On the other hand, subjective ratings do provide

valuable information, and growing evidence suggests that for most purposes this kind of self-report is a viable measurement instrument [12]. At present, a multi-modal approach as employed in this study, combining subjective self-report with more objective measurements such as physiological parameters, seems to be the most promising approach [37].

4.3. Conclusions

This study aims to contribute to the growing literature on affective computing [51] and on affect in HCI [2]. With advanced interfaces and computerized services becoming more and more available [61], it is of major importance to better understand the impact on users' affect and emotion. We focused on basic affective responses of users when experiencing ordinary system messages such as input requests, status notifications, and error messages. Employing a multi-modal approach, measuring subjective ratings as well as physiological parameters commonly associated with affective responses, it could be established that system messages trigger detectable changes in users' affective responses.

Moreover, the mode in which message information is conveyed does influence affective responses. In particular, we may conclude that presenting information as spoken messages, in contrast to standard textual messages, will shift affective responses towards more pleasant and more dominant feelings, especially concerning input and status messages. Designers should try to use the human voice to convey information when appropriate. It might be beneficial to exchange simple sound notifications such as beeps with voice notifications, which can be very short but friendly. In particular with respect to error messages and warnings we suggest to use speech notification, as this might, if carefully designed, make the user feel more dominant and in control.

System messages are a special kind of interruptions, which have for the most part been studied in terms of increasing mental load and its detrimental effects on attention and

performance [87, 88]. Bailey and Konstan [3] are among the few studies demonstrating affective effects, that is, negative feelings, of interruptions. However, affective processes may also themselves contribute to an increase in mental load, and from our findings we would advocate to address this issue more closely [7].

Mapping particular instantiations of system messages into an affective space constituted by basic dimensions such as valence, dominance, and arousal, and possibly supplemented by physiological components, might allow interface designers to choose appropriate modalities and formats to convey system messages in a more optimal way, aiming at eliciting pleasant, or at least avoiding unpleasant, affective responses in users. Knowledge about feasible ways to improve user satisfaction by tailoring system messages in an affectively suitable way will, we presume, make up one important component of a future full-fledged affective computing environment.

References

[1] R.D. Ward, P.H. Marsden, Affective computing: problems, reactions and intentions, Interacting with Computers, 16 (2004) 707-713.

[2] C. Peter, R. Beale, Affect and emotion in human-computer interaction. Lecture Notes in Computer Science, in: LNCS, Springer, Berlin, 2008.

[3] B.P. Bailey, J.A. Konstan, On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state, Computers in Human Behavior, 22 (2006) 685-708.

[4] C. Speier, I. Vessey, J.S. Valacich, The effects of interruptions, task complexity, and information presentation on computer-supported decision making performance, Decision Sciences, 34 (2003) 771-797.

[5] C.S. Carver, Approach, avoidance, and self-regulation of affect and action, Motivation & Emotion, 30 (2006) 105-110.

[6] D. Keltner, J.S. Lerner, Emotion, in: D.T. Gilbert, S.T. Fiske, G. Lindzey (Eds.) The handbook of social psychology, Wiley, New York, 2010, pp. 317-352.

[7] J. Yiend, The effects of emotion on attention: A review of attentional processing of emotional information, Cognition & Emotion, 24 (2010) 3-47.

[8] T. Partala, V. Surakka, The effects of affective interventions in human-computer interaction, Interacting with Computers, 16 (2004) 295-309.

[9] I.C. Christie, B.H. Friedman, Autonomic specificity of discrete emotion and dimensions of affective space: A multivariate approach, International Journal of Psychophysiology, 51 (2004) 143-153.

[10] S.H. Fairclough, L. Venables, Prediction of subjective states from psychophysiology: A multivariate approach, Biological Psychology, 71 (2006) 100-110.

[11] M.M. Bradley, P.J. Lang, Measuring emotion: Behavior, feeling, and physiology, in: R.D. Lane, L. Nadel (Eds.) Cognitive neuroscience of emotion, Oxford University Press, New York, 2000, pp. 242-276.

[12] M. Isomursu, M. Tähti, S. Väinämö, K. Kuutti, Experimental evaluation of five methods for collecting emotions in field settings with mobile applications, International Journal of Human-Computer Studies, 65 (2007) 404-418.

[13] M. Davis, P.J. Lang, Emotion, in: M. Gallagher, R.J. Nelson (Eds.) Handbook of psychology: Biological perspectives, Wiley, Hoboken, NJ, 2003, pp. 405-439.

[14] D. Moffat, Personality parameters and programs, in: R. Trappl, P. Petta (Eds.) Creating Personalities for Synthetic Actors, Springer, Berlin, 1997, pp. 120-165.

[15] L.F. Barrett, Valence as a basic building block of emotional life, Journal of Research in Personality, 40 (2006) 35-55.

[16] M. Cabanac, Pleasure: The common currency, Journal of Theoretical Biology, 155 (1992) 173-200.

[17] J.A. Russell, Core affect and the psychological construction of emotion, Psychological Review, 110 (2003) 145-172.

[18] A. Ortony, G.L. Clore, A. Collins, The cognitive structure of emotions, Cambridge University Press, Cambridge, MA, 1988.

[19] J. Forgas, Mood and judgment: The affect infusion model (AIM), Psychological Bulletin, 117 (1995) 39-66.

[20] R.B. Zajonc, On the primacy of affect, American Psychologist, 39 (1984) 117-123.

[21] A.M. Isen, Positive affect, in: T. Dalgleish, M.J. Power (Eds.) Handbook of cognition and emotion, Wiley, Chichester, 1999, pp. 521-539.

[22] P.J. Lang, M.M. Bradley, B.N. Cuthbert, Emotion, motivation, and anxiety: Brain mechanisms and psychophysiology, Biological Psychiatry, 44 (1998) 1248-1263.

[23] J.T. Larsen, A.P. McGraw, B.A. Mellers, J.T. Cacioppo, The agony of victory and thrill of defeat: Mixed emotional reactions to disappointing wins and relieving losses, Psychological Science, 15 (2004) 325-330.

[24] J.A. Russell, A. Mehrabian, Evidence for a three-factor theory of emotion, Journal of Research in Personality, 11 (1977) 273-294.

[25] C.E. Osgood, G.J. Suci, P.H. Tannenbaum, The measurement of meaning, University of Illinois Press, Urbana, IL, 1957.

[26] A. Bandura, Self-efficacy. The exercise of control, Freeman, New York, 1997.

[27] W. James, What is an emotion?, Mind, 9 (1884) 188-205.

[28] R.W. Levenson, Autonomic specificity and emotion, in: R.J. Davidson, K.R. Scherer, H.H. Goldsmith (Eds.) Handbook of affective sciences, Oxford University Press, New York, 2003, pp. 212-224.

[29] J.T. Cacioppo, G.G. Berntson, J.T. Larsen, K.M. Poehlmann, T.A. Ito, The psychophysiology of emotion, in: R. Lewis, J.M. Haviland-Jones (Eds.) Handbook of emotions, Guilford, New York, 2000, pp. 173-191.

[30] I.B. Mauss, R.W. Levenson, L. McCarter, F.H. Wilhelm, J.J. Gross, The tie that binds? Coherence among emotion experience, behavior, and physiology, Emotion, 5 (2005) 175-190.

[31] C. Peter, A. Herbon, Emotion representation and physiology assignments in digital systems, Interacting with Computers, 18 (2006) 139-170.

[32] E.L. van den Broek, V. Lisy, J.H. Janssen, J.H.D.M. Westerink, M.H. Schut, K. Tuinenbreijer, Affective man-machine interface: Unveiling human emotions through biosignals, in: A. Fred, J. Filipe, H. Gamboa (Eds.) Biomedical engineering systems and techologies. Communications in computer and information science, Springer, Berlin, 2010, pp. 21-47.

[33] C.M. Van Reekum, T. Johnstone, R. Banse, A. Etter, T. Wehrle, K.R. Scherer, Psychophysiological responses to appraisal dimensions in a computer game, Cognition and Emotion, 18 (2004) 663-688.

[34] J. Tao, T. Tan, Affective information processing, in, Springer, London, 2009.

[35] R.D. Ward, P.H. Marsden, Physiological responses to different WEB page designs, International Journal of Human-Computer Studies, 59 (2003) 199-212.

[36] S. Brave, C. Nass, Emotion in human-computer interaction, in: J.A. Jacko, A. Sears (Eds.) The human-computer interaction handbook, Erlbaum, Mahwah, NJ, 2002, pp. 81-96.

[37] R.L. Hazlett, J. Benedek, Measuring emotional valence to understand the user's experience of software, International Journal of Human-Computer Studies, 65 (2007) 306-314.

[38] E. Hudlicka, To feel or not to feel: The role of affect in human-computer interaction, International Journal of Human-Computer Studies, 59 (2003) 1-32.

[39] M.D. McNeese, New visions of human-computer interaction: Making affect compute, International Journal of Human-Computer Studies, 59 (2003) 33-53.

[40] R.W. Picard, Affective computing, The MIT Press, Cambridge, MA, 1997.

[41] B. Reeves, C. Nass, The media equation: How people treat computers, television, and new media like real people and places, Cambridge University Press, Stanford, CA, 1996.

[42] M.G. Helander, M.P. Tham, Hedonomics - affective human factors design, Ergonomics, 46 (2003) 1269-1272.

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

[44] S. Mahlke, M. Thüring, Studying antecedents of emotional experiences in interactive contexts., in: CHI 2007 Proceedings, ACM Press, New York, 2007, pp. 915-918.

[45] S.J.d.R. Baker, S.K. D'Mello, M.M.T. Rodrigo, A.C. Graesser, Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments, International Journal of Human-Computer Studies, 68 (2010) 223-241.

[46] R.E. Ferdig, P. Mishra, Emotional responses to computers: Experiences in unfairness, anger, and spite, Journal of Educational Multimedia and Hypermedia, 13 (2004) 143-161.

[47] C. von Scheve, R. von Luede, Emotion and social structures: Towards an interdisciplinary approach, Journal for the Theory of Social Behaviour, 35 (2005) 303-328.

[48] W.F. Bridsall, Web 2.0 as a social movement, in: Webology, 2007.

[49] H.-R. Pfister, M. Oehl, The impact of goal focus, task type and group size on synchronous net-based collaborative learning discourses, Journal of Computer Assisted Learning, 25 (2009) 161-176.

[50] H.-R. Pfister, M. Mühlpfordt, W. Müller, Lernprotokollunterstütztes Lernen - ein Vergleich zwischen unstrukturiertem und systemkontrolliertem diskursivem Lernen im Netz [Learning with learning protocols - a comparison between unstructured and system-controlled net-based discoursive learning], Zeitschrift für Psychologie - Journal of Psychology, 211 (2003) 98-109.

[51] A. Paiva, R. Prada, R.W. Picard, Affective computing and intelligent interaction, in, Springer, Berlin, 2007.

[52] R.W. Picard, J. Klein, Computers that recognize and respond to user emotion: Theoretical and practical implications, Interacting with Computers, 14 (2002) 141-169.

[53] S.H. Fairclough, Fundamentals of physiological computing, Interacting with Computers, 21 (2009) 133-145.

[54] Z. Zeng, M. Pantic, G.I. Roisman, T.S. Huang, A survey of affect recognition methods: Audio, visual, and spontaneous expressions, IEEE Transactions on Pattern Analysis and Machine Intelligence, 31 (2009) 39-58.

[55] J. Allanson, S.H. Fairclough, A research agenda for physiological computing, Interacting with Computers, 16 (2004) 857-878.

[56] B. Fasel, J. Luettin, Automatic facial expression analysis: A survey, Pattern Recognition, 36 (2003) 259-275.

[57] M. Pantic, I. Patras, Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences, IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics, 36 (2006) 433-449.

[58] R.D. Ward, B. Cahill, P.H. Marsden, C.A. Johnson, Physiological responses to HCI events - what produces them and how detectable are they?, in: Proceedings of HCI 2002, 2002, pp. 90-93.

[59] J. Klein, Y. Moon, R.W. Picard, This computer responds to user frustration: Theory, design, and results, Interacting with Computers, 14 (2002) 119-140.

[60] C. Bartneck, M.J. Lyons, Facial expression analysis, modeling, and synthesis: Overcoming the limitations of Artificial Intelligence with the art of the soluble, in: J. Vallverdu, D. Casacuberta (Eds.) Handbook of research on synthetic emotions and sociable robotics: New applications in affective computing and Artificial Intelligence, IGI Global, Hershey, PA, 2009, pp. 33-53.

[61] T. Zhang, D.B. Kaber, B. Zhu, M. Swangnetr, P. Mosaly, L. Hodge, Service robot feature design effects on user perceptions and emotional responses, Intelligent Service Robotics, 3 (2010) 73-88.

[62] R.L. Mandryk, K.M. Inkpen, T.W. Calvert, Using psychophysiological techniques to measure user experience with entertainment technologies, Behaviour & Information Technology, 25 (2006) 141-158.

[63] T. Lin, A. Imamiya, X. Mao, Using multiple data sources to get closer insights into user cost and task performance, Interacting with Computers, 20 (2008) 364-374.

[64] C. Speier, J.S. Valacich, I. Vessey, The influence of task interruption on individual decision making: An information overload perspective, Decision Sciences, 30 (1999) 337-360.

[65] L. Qiu, I. Benbasat, An investigation into the effects of text-to-speech voice and 3D avatars on the perception of presence and flow of live help in electronic commerce, ACM Transactions on Computer-Human Interaction, 12 (2005) 329-355.

[66] C. Frauenberger, T. Stockman, Auditory display design - An investigation of a design pattern approach, International Journal of Human-Computer Studies, 67 (2009) 907-922.

[67] V. Maffiolo, N. Chateau, The emotional quality of speech in voice services, Ergonomics, 46 (2003) 1375-1385.

[68] K. Kallinen, N. Ravaja, Effects of the rate of computer-mediated speech on emotion-related subjective and phsysiological responses, Behaviour & Information Technology, 24 (2005) 365-373.

[69] T. Hess, M. Fuller, D. Campbell, Designing interfaces with social presence: Using vividness and extraversion to create social recommendations agents, Journal of the Association for Information Systems, 10 (2009) 889-919.

[70] A.R. Dennis, R.M. Fuller, J.S. Valacich, Media, tasks, and communication processes: A theory of media synchronicity, MIS Quarterly, 32 (2008) 575-600.

[71] M.M. Bradley, P.J. Lang, Measuring emotions: The self-assessment manikin and the semantic differential, Journal of Behavior Therapy and Experimental Psychiatry, 25 (1994) 49-59.

[72] J.L. Andreassi, Psychophysiology: Human behavior and physiological response, 5th ed., Erlbaum, Mahwah, NJ, 2007.

[73] C. Peter, E. Ebert, H. Beikirch, A wearable multi-sensor system for mobile acquisition of emotion-related physiological data, in: J. Tao, T. Tan, R.W. Picard (Eds.) Proceedings of the 1st International Conference on Affective Computing and Intelligent Interaction ACII 2005, Springer, Berlin, 2005, pp. 691-698.

[74] A. Barreto, J. Zhai, M. Adjouadi, Non-intrusive physiological monitoring for automated stress detection in human-computer interaction, in: M. Lew, N. Sebe, T.S. Huang, E.M. Bakker (Eds.) Proceedings of HCI 2007. Lecture Notes in Computer Science, Springer, Heidelberg, 2007, pp. 29-38.

[75] Y. Gao, A. Barreto, M. Adjouadi, Comparative analysis of noninvasively monitored biosignals for affective assessment of a computer user, in: A. McGoron, C. Li, W.-C. Lin

(Eds.) 25th Southern Biomedical Engineering Conference 2009, IFMBE Proceedings 24, Springer, Berlin, 2009, pp. 255-260.

[76] W. Boucsein, The use of psychophysiology for evaluating stress-strain processes in human-computer interaction, in: R.W. Backs, W. Boucsein (Eds.) Engineering psychophysiology: Issues and applications, Erlbaum, Mahwah, NJ, 2000, pp. 289-309.

[77] M.E. Dawson, A.M. Schell, D.L. Filion, The electrodermal system, in: J.T. Cacioppo, L.G. Tassinary, G.G. Berntson (Eds.) Handbook of psychophysiology, Cambridge University Press, Cambridge, US, 2007, pp. 159-181.

[78] A. Keil, J.F. Smith, Mitchell, T. R., & Beach, L. R., B.C. Wangelin, D. Sabatinelli, M.M. Bradley, P.J. Lang, Electrocortical and electrodermal responses covary as a function of emotional arousal: A single-trial analysis, Psychophysiology, 45 (2008) 516-523.

[79] J. Fahrenberg, C.J.E. Wientjes, Recording methods in applied environments, in: R.W. Backs, W. Boucsein (Eds.) Engineering psychophysiology: Issues and applications, Erlbaum, Mahwah, NJ, 2002, pp. 111-132.

[80] P.H. Venables, M.J. Christie, Electrodermal activity, in: I. Martin, P.H. Venables (Eds.) techniques in psychophysiology, Wiley, New York, 1980, pp. 4-67.

[81] K.R. Scherer, T. Johnstone, G. Klasmeyer, Vocal expression of emotion, in: R.J. Davidson, K.R. Scherer, H.H. Goldsmith (Eds.) Handbook of affective science, Oxford University Press, New York, 2003, pp. 433-456.

[82] B.G. Tabachnick, L.S. Fidell, Using multivariate statistics, 5th ed., Pearson, Boston, 2007.

[83] A. Agresti, Categorical data analysis, 2nd ed., Wiley, New York, 2002.

[84] J.C. Pinheiro, D.M. Bates, Mixed-effects models in S and S-PLUS, Springer, New York, 2000.

[85] S. Grandhi, Q. Jones, Technology-mediated interruption management, International Journal of Human-Computer Studies, 68 (2010) 288-306.

[86] D. Zillman, P. Vorderer, Media entertainment: The psychology of its appeal, in, Erlbaum, Mahwah, NJ, 2000.

[87] L. Dabbish, R. Kraut, Awareness displays and social motivation for coordinating communication, Information Systems Research, 19 (2008) 221-238.

[88] B. Xie, G. Salvendy, Review and reappraisal of modelling and predicting mental workload in single- and multi-task environments, Work & Stress, 14 (2000) 74-99.

Tables

Table 1.

Means and Standard Deviations (in Parentheses) of Dependent Variables.

Table 2.

Summary of $2 \times 3 \times 3$ ANOVA Analysis: Cells show F-value with Degrees of Freedom (in Parentheses), and p-value.

Table 3.

Correlations Among Dependent Variables.

Figures

Figure 1. Messages used depending on global modality (written or spoken), message category (input, status, error), and specific modality; column two shows icons of message-plus-symbol condition.

Figure 2. Sequence of tasks, instructions, and measurements taken across one experimental trial.

Figure 3. Means of physiological measures as a function of global modality (written vs. spoken) and message category (input, status, error). Error bars indicate standard errors.

Figure 4. Means of subjective ratings as a function of global modality (written vs. spoken) and message category (input, status, error). Error bars indicate standard errors.

Figure1

category \ modality		written (or spoken)	written (or spoken) + symbol	written (or spoken) + sound
input request	spreadsheet calculation	Do you want to save changes?	 Do you want to save changes?	Do you want to save changes?
	information search	More than ten hits. Do you want to view the entire list?	 More than ten hits. Do you want to view the entire list?	More than ten hits. Do you want to view the entire list?
	keyboard typing training	Please choose a Username.	 Please choose a Username.	Please choose a Username.
status notification	spreadsheet calculation	With the changes you are about to make your travel costs will be more than doubled.	 With the changes you are about to make your travel costs will be more than doubled.	With the changes you are about to make your travel costs will be more than doubled.
	information search	No matches could be found.	 No matches could be found.	No matches could be found.
	keyboard typing training	The tutorial is finished now.	 The tutorial is finished now.	The tutorial is finished now.
error message	spreadsheet calculation	Changes in this field are just possible by using the formula manager.	 Changes in this field are just possible by using the formula manager.	Changes in this field are just possible by using the formula manager.
	information search	Check your input. Please tip ‘Surname, first Name’.	 Check your input. Please tip ‘Surname, first Name’.	Check your input. Please tip ‘Surname, first Name’.
	keyboard typing training	You made a spelling mistake.	 You made a spelling mistake.	You made a spelling mistake.

Figure2

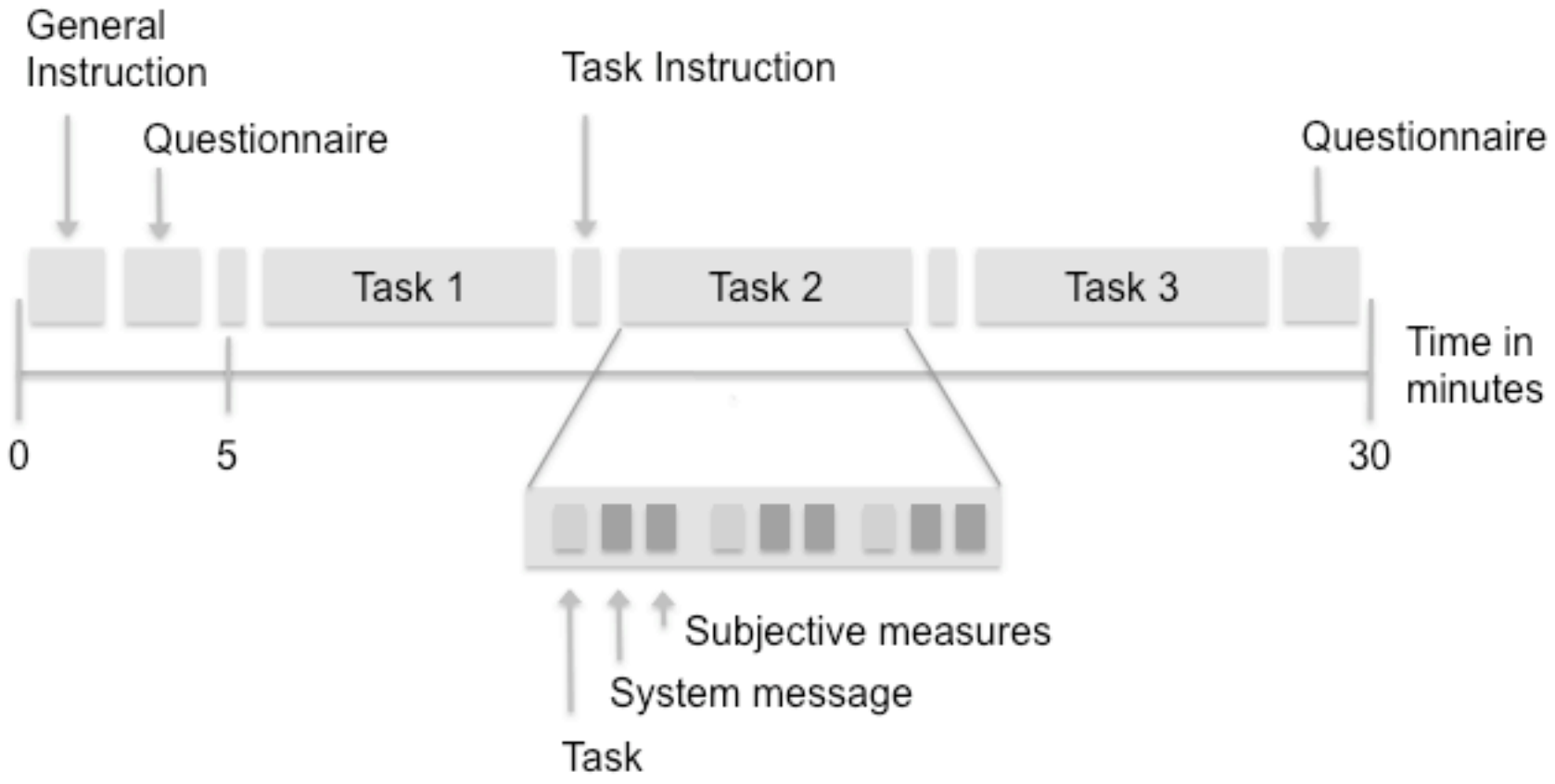
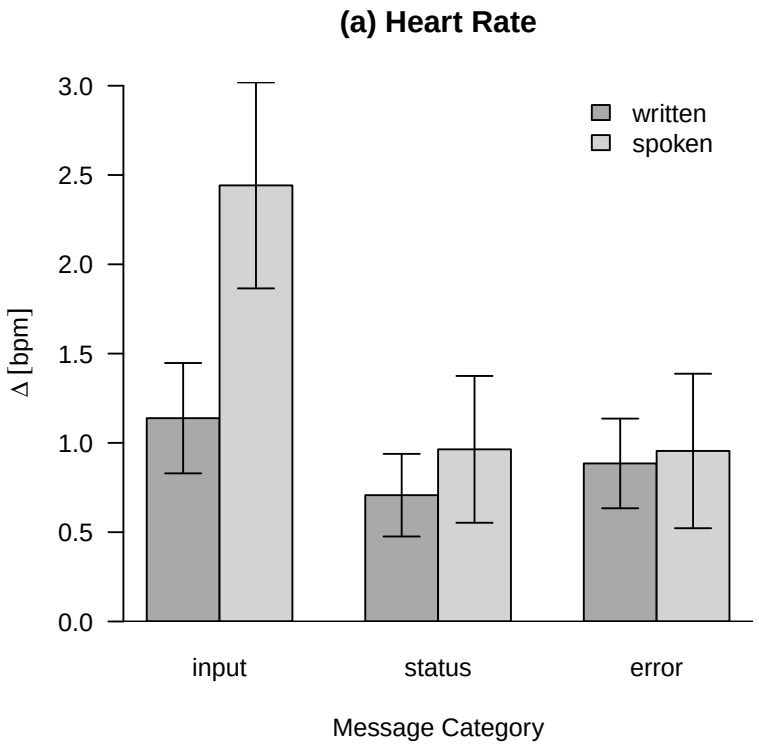
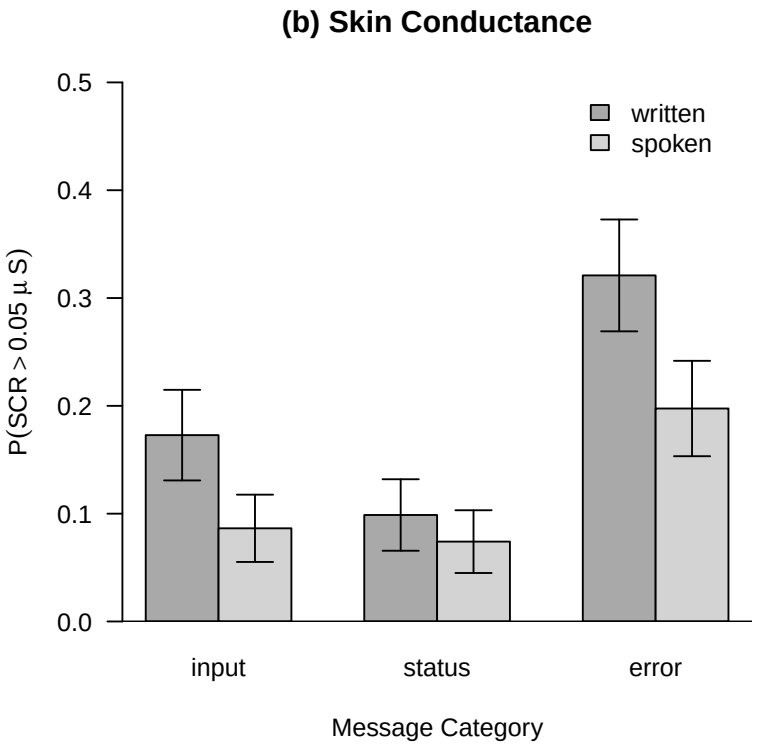

General Instruction
Questionnaire
Task Instruction
Questionnaire
Task 1
Task 2
Task 3
Time in minutes
0
5
30
Subjective measures
System message
Task

Figure3


(a) Heart Rate
written
spoken
Δ [bpm]
3.0
2.5
2.0
1.5
1.0
0.5
0.0
input
status
error
Message Category


(b) Skin Conductance
written
spoken
P(SCR > 0.05 μ S)
0.5
0.4
0.3
0.2
0.1
0.0
input
status
error
Message Category

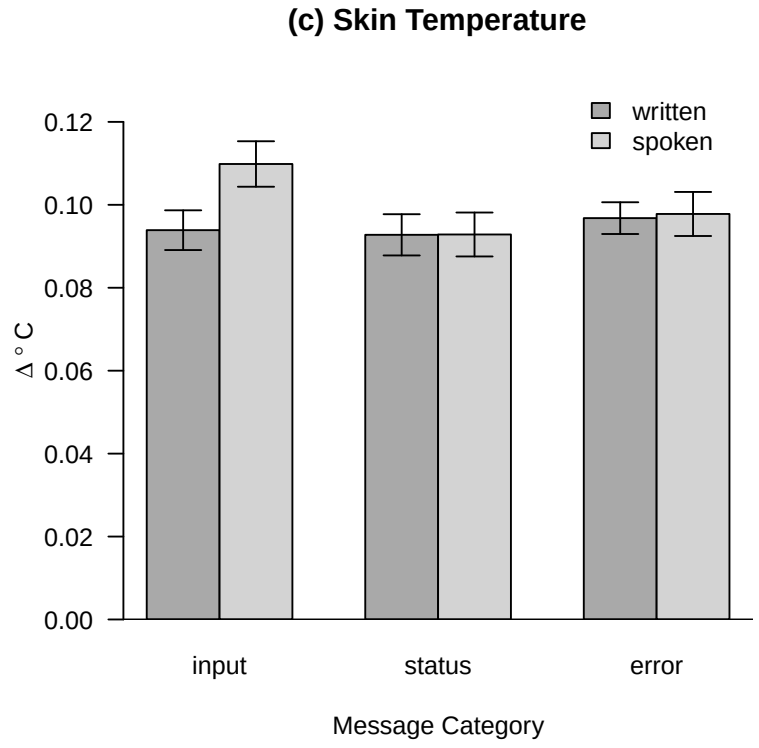

(c) Skin Temperature
written
spoken
Δ ° C
0.12
0.10
0.08
0.06
0.04
0.02
0.00
input
status
error
Message Category

Figure4

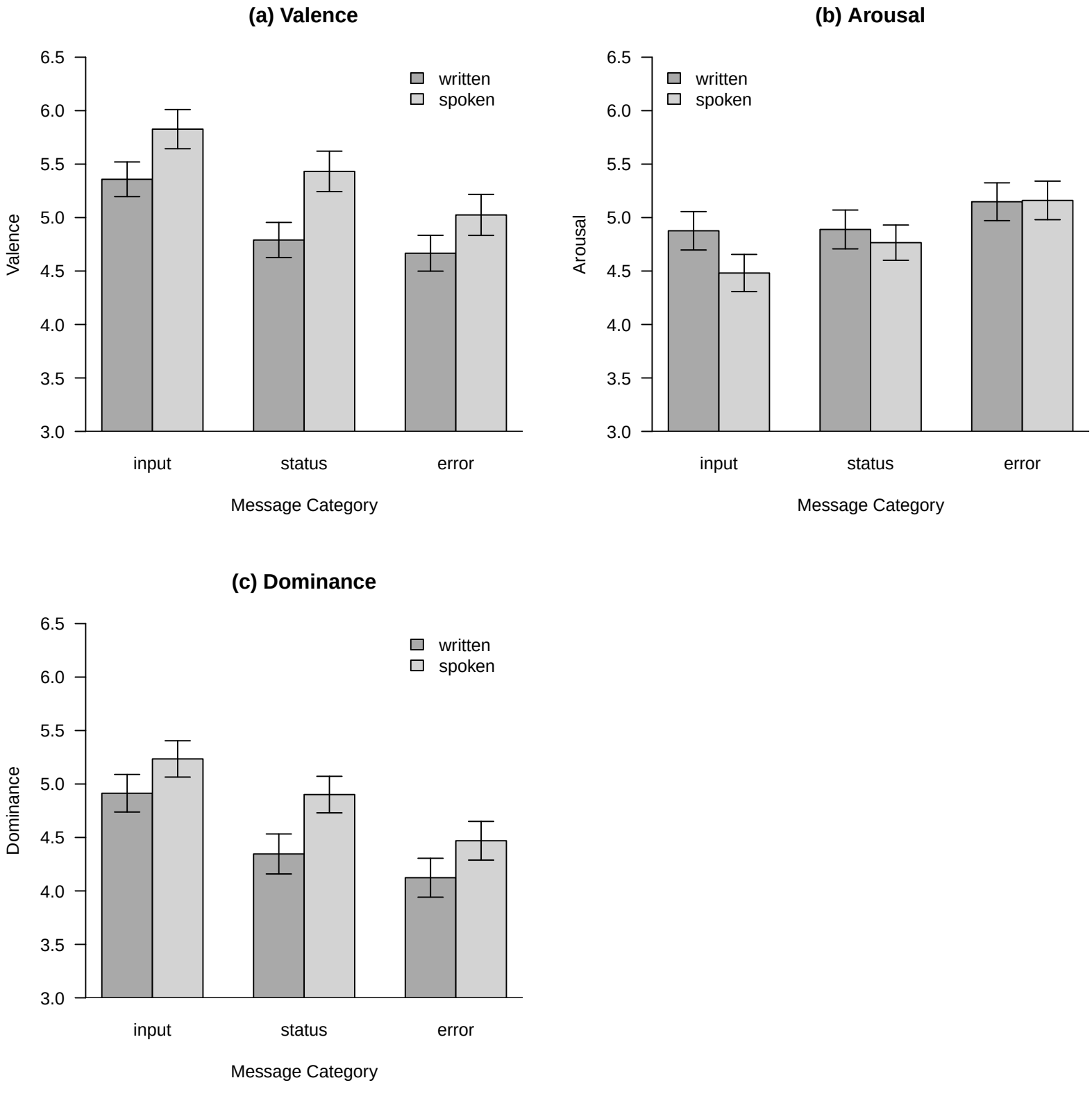

(a) Valence
Valence
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
written
spoken
input
status
error
Message Category
(b) Arousal
Arousal
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
written
spoken
input
status
error
Message Category
(c) Dominance
Dominance
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
written
spoken
input
status
error
Message Category

Table1

Table 1.

Means and Standard Deviations (in Parentheses) of Dependent Variables.

Measure	Written			Spoken		
	Input	Status	Error	Input	Status	Error
HR [bpm]	1.138 (2.781)	0.707 (2.081)	0.885 (2.260)	2.442 (5.190)	0.964 (3.699)	0.955 (3.893)
SC [μS]	0.018 (0.043)	0.018 (0.065)	0.039 (0.071)	0.013 (0.047)	0.019 (0.068)	0.037 (0.090)
TM [°C]	0.094 (0.043)	0.093 (0.045)	0.097 (0.034)	0.110 (0.049)	0.093 (0.048)	0.098 (0.048)
Valence	5.36 (1.46)	4.79 (1.48)	4.67 (1.51)	5.83 (1.65)	5.43 (1.70)	5.02 (1.72)
Arousal	4.88 (1.62)	4.89 (1.64)	5.15 (1.59)	4.48 (1.57)	4.77 (1.49)	5.16 (1.62)
Dominance	4.91 (1.58)	4.35 (1.68)	4.12 (1.649	5.23 (1.58)	4.90 (1.54)	4.47 (1.63)

Note. N = 27 for written and spoken global modality, respectively. HR (heart rate), SC (skin conductance response), and TM (temperature) denote phasic values (differences). Valence, arousal, and dominance are measured on a 9-point scale.

Table2

Table 2.

Summary of $2 \times 3 \times 3$ ANOVA Analysis: Cells show F-value with Degrees of Freedom (in Parentheses), and p-value.

	HR	SC	TM	Valence	Arousal	Dominance
Global Modality	1.92 (1,50) p = .172	0.032 (1,50) p = .856	0.67 (1,50) p = .417	5.15 (1,50) p = .028*	0.33 (1,50) p = .569	2.49 (1,50) p = .121
GM × SM	1.55 (2,104) p = 0.216	2.20 (2,104) p = .116	0.34 (2,104) p = .710	0.05 (2,104) p = .949	0.80 (2,104) p = .452	0.42 (2,104) p = .660
GM × MC	2.21 (2,104) p = 0.115	0.21 (2,104) p = .815	2.26 (2,104) p = .109	0.42 (2,104) p = .658	1.05 (2,104) p = .355	0.47 (2,104) p = .629
Specific Modality	0.35 (2,104) p = 0.708	1.49 (2,104) p = .229	1.33 (2,104) p = .268	0.55 (2,104) p = .576	0.05 (2,104) p = .950	1.72 (2,104) p = .184
SM × MC	0.38 (4,208) p = 0.819	0.56 (4,208) p = .690	1.86 (4,208) p = .119	0.48 (4,208) p = .754	0.70 (4,208) p = .594	0.48 (4,208) p = .747
Message Category	5.57 (2,104) p = 0.005*	9.42 (2,104) p < .001*	2.34 (2,104) p = .102	11.8 (2,104) p < .001*	5.75 (2,104) p = .004*	17.1 (2,104) p < .001*
GM × SM × MC	0.41 (4,208) p = .798	1.47 (4,208) p = .211	1.79 (4,208) p = .131	1.22 (4,208) p = .302	0.26 (4,208) p = .902	1.61 (4,208) p = .173

Note. * indicates significant effect ($\alpha = 5\%$). × indicates interaction effects (GM = Global Modality, SM = Specific Modality, MC = Message Category).

Table3

Table 3.

Correlations among Affective Response Measures.

	valence	arousal	dominance	heart rate	skin conductivity
arousal	-.59**				
dominance	.67**	-.55**			
heart rate	-.01	-.08	.06		
skin conductivity	-.18	.19	.01	.07	
temperature	.10	-.14	.08	-.02	-.05

Note. ** indicates correlations significant at $\alpha = .01$; N = 54.

Revision of manuscript IwC 2589 "Affective responses to system messages in Human-Computer-Interaction: Effects of modality and message type".

Responses to Reviewer Comments – Revision 2

Reviewer 1

1. Justification of the use of ANOVA on ratings data (SAM).

In the results section on p. 18 we include a paragraph (2nd paragraph) to detail the analysis strategy for the ratings data. The assumptions of ANOVA were largely met (normality, homogeneity of variance for between-subjects factors); sphericity for repeated factors was not always met but applying a Greenhouse-Geisser correction did not change any statistical conclusions. Furthermore, we conducted parallel analyses treating the rating scales as strictly ordinal (proportional odds regression model), but obtained virtually identical results as from ANOVA; thus we decided to report only ANOVA results.

2. Justification of physiological measures

We elaborated the respective paragraph (para 2 on page 14) emphasizing the need for unobtrusive measurement.

3. Use of logistic regression on the SCR data.

On page 19, 2nd paragraph, we now justify the dichotomization of the SCR variable and the then appropriate use of logistic regression more explicitly. Still we thought it informative to also present the results of an ANOVA, to be consistent with the overall analysis strategy and to keep results comparable for the reader.

Reviewer 1 also advised to replace and add a reference concerning physiological evaluation of computer games and addresses pacing and reward, which we did on p. 9 (1st paragraph, reference [62]).

Reviewer 2

- p. 15: details on the equipment are now added in the respective paragraphs on pp. 14/15 and a new reference about the Fraunhofer system is added (Peter et al. 2005).
- p. 17: corrected
- p. 20: corrected