Shared Experiences of Technology and Trust: An Experimental Study of Physiological Compliance Between Active and Passive Users in Technology-Mediated Collaborative Encounters

Enid Montague, Jie Xu, and Erin Chiou

***Abstract*—The aim of this study is to examine the utility of physiological compliance (PC) to understand shared experience in a multiuser technological environment involving active and passive users. Common ground is critical for effective collaboration and important for multiuser technological systems that include passive users since this kind of user typically does not have control over the technology being used. An experiment was conducted with 48 participants who worked in two-person groups in a multitask environment under varied task and technology conditions. Indicators of PC were measured from participants' cardiovascular and electrodermal activities. The relationship between these PC indicators and collaboration outcomes, such as performance and subjective perception of the system, was explored. Results indicate that PC is related to group performance after controlling for task/technology conditions. PC is also correlated with shared perceptions of trust in technology among group members. PC is a useful tool for monitoring group processes and, thus, can be valuable for the design of collaborative systems. This study has implications for understanding effective collaboration.**

***Index Terms*—Group performance, multiagent systems, passive user, physiological compliance (PC), trust in technology.**

I. Introduction

A CUSTOMER and a service representative finalize a transaction using a computer. A pilot and a copilot fly an airplane with an autopilot system. A doctor and a patient share a computer monitor to discuss treatment options. These vignettes describe joint-activity systems, which are as common as people interacting with or through technology.

Effective joint activity requires common ground among the group members [1] or a shared understanding of the relevant knowledge, beliefs, and assumptions that support cooperative coordination, and the process by which this shared understanding is updated [2]. Common ground includes various types of understanding, such as understanding each person's role, skills, goals, or current state [2]. It also includes an awareness of the environment, the tools being used, and the tasks being performed [3]. With common ground, agents of joint activities can understand potentially ambiguous communication confidently and efficiently. Common ground also facilitates an agent's anticipation of future states of the activity and the ability to respond to changes appropriately. Thus, common ground is considered critical for joint activity and coordination among multiple agents.

Understanding common ground and how groups leverage this information to coordinate an activity can be important for understanding successful joint activities in complex and dynamic environments. Joint activities also occur between people and technology, which is why members of a joint activity might be referred to as "agents," or between people with technology as a communication medium. Particularly as workplaces move farther apart geographically and virtual teams become more common, computers and information technology become the carriers of information that update the common ground and status of the activity. For example in healthcare, health information technology is used to coordinate patient information [4], [5]. However, it is shown that introducing a new technology such as the electronic health records may have unintended consequences to the care process [6]. In a primary care health encounter, a doctor's computer monitor-sharing with a patient has been described as an essential element of patient-centered care because it improves patient education and the shared decision-making process [7]. In addition, monitor-sharing might help create common ground between patients and their doctors by facilitating communication and discussion during the visit [8].

Common ground builds on prior knowledge before an activity, history of the activity, and the current state of the activity [2]. Maintaining common ground among team members might be achieved in three ways: communicating directly, monitoring shared displays, and monitoring their environment (including monitoring each other's behavior) [9]. In a complex and dynamic environment with a large amount of uncertainty, updating the common ground is especially important. In a multiuser system involving passive users, such as the doctor–patient example where patients are not usually in direct control of the technology, communication about the current state can be the most important factor for building common ground between passive and active users [10]. This is also the case for joint users with different levels of expertise or perspectives, such as doctors and nurses, or service personnel and customers [11].

Manuscript received August 23, 2013; revised March 25, 2014 and December 30, 2013; accepted October 14, 2013. Date of publication June 17, 2014; date of current version September 12, 2014. This work was supported in part by the Air Force Office of Sponsored Research Trust and Influence Program. This paper was recommended by Associate Editor J. Wachs.

E. Montague is with the Division of General Internal Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611 USA (e-mail: enid.montague@northwestern.edu).

J. Xu and E. Chiou are with the Department of Industrial and Systems Engineering, University of Wisconsin-Madison, Madison, WI 53705 USA (e-mail: jxu29@wisc.edu; chiou2@wisc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2014.2325859

One challenge to understanding common ground is measuring it dynamically in teams, using methods beyond self-report or observation. Finally, little is known about how achieving common ground affects individuals on the team.

II. Physiological Compliance

One potential approach to inferring the real-time cognitive state of individuals during joint activity is to use physiological measures [2], [12]. Comparing real-time physiological measures, such as those linked to concurrent tasks or reactions of team members, may provide a way to assess the status of common ground during joint activity. Physiological measures could complement other approaches, such as observation and subjective reports, with the advantages of being unobtrusive [13], objective, and providing information in real time [14].

Physiological measures have been used to identify response patterns that relate to affective exchange among individuals in a joint activity during direct communication. For example, Levenson and Gottman [15] found that physiology correlates between a wife and a husband during a conversation could predict the marital satisfaction rating of the couple. Another study [16] found that covariation in electrodermal activities of two individuals relates to conversations with positive or negative affective behaviors but not the neutral ones. Distinctive physiological changes can also be observed when an individual monitors another individual's behavior, one of the ways to update common ground. For example, Cwir *et al*. [17] found that in a two-person group, if one person observes another person running, the observer will show physiological arousal similar to that of the runner. A study by Nowlin *et al*. [18] found similar results in social interactions; passive participants showed similar physiological changes with the active participants.

One useful concept in investigating shared physiological change among multiple individuals is physiological compliance (PC). PC is the study of covariation in participants' physiological changes during social interaction [19]. Many studies about PC link it to group performance and effectiveness. For example, higher PC was found to relate to shorter task completion time in a collaborative tracking task [20] and better performance in a military building clearing task [21].

There are several PC indicators used to determine the PC of two individuals, such as signal matching (SM), instantaneous derivative matching (IDM), directional agreement (DA), cross correlation (CC), and weighted coherence (WC). SM, IDM, and DA were developed by Elkins *et al*. [21]. CC and WC were used by Henning *et al*. [20] to measure PC; WC was developed based on Porges *et al*.'s work [20], [22]. In SM, the differences between the areas of the two individuals' data curves are compared. The smaller the area between the curves, the lower the SM score. A low SM score indicates higher curve similarity and thus higher PC. In IDM, the slopes of the curves are compared. This is accomplished by averaging the differences between the instantaneous derivatives of corresponding points between the curves; the derivative of a point provides the tangent and thus the slope of the curve at that point. A low IDM score would indicate high similarity between the curves, and thus high PC. In the third PC measure, DA, the directional movement of the curve is assessed by comparing each point on the curve with its previous point. For example, an increasing directional movement would mean data point B is higher in value than data point A. A percentage of the two curves' DA is then calculated; a higher percentage indicates higher PC. A CC coefficient is calculated to determine the covariance of corresponding data points in each physiological data curve at lag 0. While the PC indicators discussed previously were derived from the original physiological signal, which recorded on the time domain, WC is a PC indicator that concerns frequency domain. WC quantifies the similarities of two individuals' physiology responses on a specified frequency band regardless of phase differences. These PC indicators are summarized in Table I.

TABLE I
PC Indicators and Their Definitions

PC indicator	Definition
SM	The difference between the areas under the curves of two physiological signals.
IDM	The average of the differences between the instantaneous derivatives of corresponding points between the curves of two physiological signals.
DA	The average agreement of the direction of change of corresponding points comparing to their previous points in the curves of two physiological signals.
CC	Lag 0 CC of the curves of two physiological signals.
WC	The similarity of two curves of physiological signals on a specified frequency band.

III. Current Study

This study aimed to understand the process of joint activity, using PC measures, in technologically complex environments that include active and passive users. This study also investigated how PC relates to task performance and the subjective experience of the group.

As the characteristics of a task or technology change, a group may need to adapt to maintain common ground. For example, they may need to communicate more frequently or monitor each other's behavior more closely. Behavioral changes would likely be reflected as PC indicators change.

Hypothesis 1: Task/technology conditions (specifically, task demand and technology reliability) affects the level of group PC. Specifically, high task demand and low technology reliability are related to high PC.

Empirical studies found that PC is related to group/team coordination and performance in dynamic task environments [20], [21]. This study tested if PC was related to task performance after controlling for the effects of varied task/technology conditions introduced by experimental manipulations.

Hypothesis 2: PC of a group is positively related to the performance of the group after controlling for effects of task/technology conditions.

Previous studies found that PC relates to affective exchange and corresponding subjective ratings such as satisfaction or how much a participant liked working with the partner [13], [15], [16]. This study investigated the relationship between PC and subjective experiences, focusing on passive users' awareness of active users' workload. Common ground includes understanding of the states of other individuals in a group, such as mutual awareness of workload [23].

Hypothesis 3: PC of a group is related to how the passive user estimates the active user's workload. Specifically, the higher the PC, the closer the passive user's estimation of the active user's workload is to the active user's self-reported workload.

PC may also relate to the shared perception of the trustworthiness of the technology being used. Trust in technology is an individual's attitude that a technology will help achieve a specific goal [24]. Trust in technology influences how the technology is used and may influence a user's trust in other elements of a work system, such as co-users and the institution which implements the technology [25], [26]. Trust in technology may be calibrated to a similar level among the users while common ground is being established.

Hypothesis 4: PC of a group is related to the shared perception about the trustworthiness of the technology among the group members. Specifically, high PC is related to high similarity of the group members' level of trust in technology.

IV. Method

A. Participants

This study recruited participants from an introduction to human factors course in a large Midwestern university in the U.S. The sample size was 48. The participants ranged in age from 19 to 29 ($\mathrm{mean} = 21.6, \mathrm{SD} = 1.7$). Fifteen of the participants were female (31.3%). The participants received extra course credits for participating in the study. The participants were randomly assigned to two-person groups ($n = 24$) for the experiment. Among the groups, 50% of the participants reported that they did not know the other member in the group, or were only acquainted for a short period of time in the course prior to the experiment; 50% of the participants reported that they knew each other prior to the experiment. In addition, 50% of the groups were mixed-gender groups where the group members had different genders. The two groups who achieved the highest performance were awarded \$20 per group member. The protocol of this study was approved by the university's institutional review board.

B. Task

The participants performed psycho-motor tasks in a modified version of the Multiattribute Task Battery (MATB) program [27] which runs on a PC-DOS operating system. The computer workstations for the participants to perform the task were equipped with 23-in monitors and keyboards, mice, and joysticks as control devices. Three tasks of the MATB were used: a monitoring task, a tracking task, and a resource management task. The monitoring task required participants to respond as quickly as possible to lights and dial fluctuations via keyboard keystrokes. The tracking task required participants to maintain a randomly moving target in the center of the screen using a joystick. The resource management task required participants to control several pumps to maintain optimum liquid levels in two tanks. The MATB also contained an audio task but it was not used in this study. The participants had to perform all three tasks at the same time, and they were instructed that the three tasks had equal importance.

C. Procedure

The participants participated in two-person groups. One of the participants was randomly assigned as the active user, and the other participant was assigned as the passive user. The active user had full access to the control devices of the computer, the keyboard, and the joystick. The passive user did not have access to any of the control devices, but he/she could monitor the tasks and communicate with the active user to assist with the task. The participants were instructed to work as a team, and if they achieved the highest performance, both of them would be eligible for the monetary award.

Upon arrival, participants were asked to provide informed consent. Then, the participants were connected to the physiological measurement device, and a 6-min baseline (BL) recording was collected. During the BL recording, the participants were instructed not to interact with each other. After the BL recording, written instructions of the MATB was given to the participants to read. After the reading period, the participants went through a 6-min training session, separately, on different computers. After the training session, the experimenter answered the participants' questions about the control and the goal of the tasks and received verbal confirmation that the participants understood the tasks. The participants then worked together on one computer workstation (with one shared display) for the task trials. There were three task trials each lasting a fixed length of 6 min. The task trials varied in difficulty level and technology reliability level, and the sequence of these conditions was counterbalanced across groups. Participants completed surveys about their subjective experience after each task trial. Scales used included the NASA task load index (NASA-TLX) [28] and a trust in technology scale [29]. The participants moved to the next task trial once both of them finished the survey. Physiological data were collected during the task trials. All the participants completed a demographic information survey at the end of the experiment. Finally, the participants were debriefed about the study. The procedure is summarized in Fig. 1.

D. Experiment Design

The within-subject experimental treatment of this study was based on the three levels of the task/technology conditions: the normal condition, the hard condition, and the low reliability condition. The participants went through these three levels in the three task trials. The sequence of the conditions was counterbalanced across groups. In the normal condition, the demand level in the monitoring task was low, the difficulty level in the tracking task was low, and the reliability of the pumps in the resource management task was high. In the hard condition, the demand level in the monitoring task was high, in that more frequent responses from the user were needed, and the tracking task was more difficult, in that the speed and movement

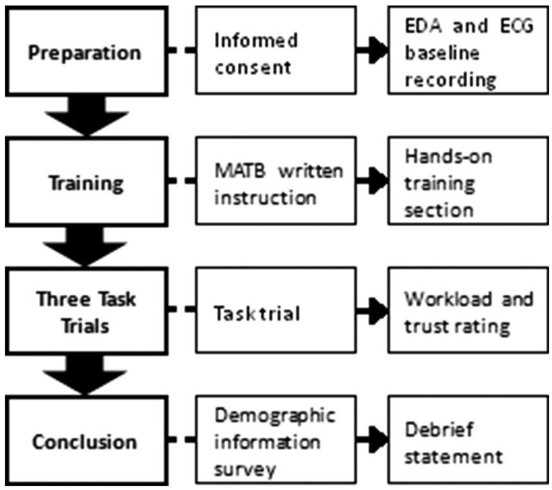

Fig. 1. Procedure of the current experiment.

TABLE II
MANIPULATION OF THE WITHIN-SUBJECT VARIABLE—TASK/TECHNOLOGY CONDITIONS

MATB tasks	Levels of task/technological conditions		
	Normal condition	Hard condition	Low reliability condition
Monitoring task	Standard demand	High demand	Standard demand
Tracking task	Standard difficulty	High difficulty	Standard difficulty
Resource management task	Reliable pumps	Reliable pumps	Unreliable pumps

amplitude of the target were increased. All the parameters of the resource management task were the same as in the normal condition. In the low reliability condition, the difficulty level in the monitoring task and the tracking task were the same as in the normal condition. However, the pumps in the resource management task became unreliable, in that certain pumps would "fail." When a pump fails, its flow rate becomes zero and turns red in color, and the participant is not able to control the flow rate. Thus, it was difficult for the participants to maintain the desired liquid level in this condition. In sum, the normal condition was the control condition for group performance, the hard condition was expected to alter the group's performance in the monitoring task and the tracking task, and the low reliability condition was expected to alter the group's performance in the resource management task. The manipulation of this variable is summarized in Table II.

E. Measurements

1) Performance: Performance in the monitoring task was measured in MATB by average reaction time, miss rate, and false alarm rate. These measurements were first standardized using z-transformation applied to the whole dataset. The values were then reversed scaled so that the higher values represented higher performance. Finally, one single performance index was calculated for the monitoring task by averaging across the three of them. Performance in the tracking task was first measured by the target's root-mean-square error (RMSE) which indicates the deviation of the target from the center of the screen. Then, the RMSEs were reversed so that a higher value would represent higher performance. Performance in the resource management task was first measured by mean absolute deviation of the liquid level of the tanks to the desired level. The mean absolute deviation values were also reversed so that higher values presented higher performance.

2) Physiological Compliance: Physiology data were collected with a MP150 Data Acquisition System (Biopac Systems Inc.) and Acqknowledge (version 4.2). The data gathered from the participants included electrodermal activity (EDA) and cardiovascular activity measurements. EDA measures the changes in the volar surface of the fingers' skin conductance level as well as skin conductance response (SCR) due to sweat gland (eccrine) activity. Two electrodes were placed on the distal phalanx area of the index and middle finger of the participants to measure EDA. For the active users, electrodes were placed on the hands they used for controlling the joystick to minimize intrusiveness of the measurement, since no key pressing was required for the operation of the joystick. Electrocardiogram (ECG) was used for measuring participants' cardiovascular activity . Three electrodes were used for measuring ECG. Two electrodes were placed below the left and right clavicle. The third electrode was placed below the left pectoral muscle. The sample rate was 62.5 Hz for EDA and 1000 Hz for the ECG. The signals from the two participants in a group were synchronized during data collection.

Low-pass filters at 5 Hz were applied to the EDA data to filter high frequency (HF) noise in the signals. The EDA signals were resampled at 2 Hz before subsequent analysis. Interbeat intervals (IBIs) were derived from the ECG. All the derived IBI data were visually examined for artifacts. If an artifact was identified in the IBI series, the corresponding original ECG data would be examined and corrected. The correction would be done by manually removing artificial peaks or recovering R waves. New IBI data would be derived again after the corrections. All the IBI data were resampled at 2 Hz. Data from the same participant were standardized using z-transformation for both EDA and IBI.

PC indicators, including SM, IDM, DA, CC, and WC, were calculated for the physiology data from each group in the baseline recording (BL) and the three task trials. PC indicators in the time domain, including SM, IDM, DA, and CC, were calculated from the standardized EDA and IBI data. For CC, the lag 0 CC coefficient was used. In addition, CC was calculated directly from the data without detrending or an autocorrelation adjustment. WC was calculated in two steps using R [30]. First, raw periodograms were derived from the data with fast Fourier transformations after linear trends were removed; and the periodograms were smoothed using modified Daniell smoothers, which are moving averages giving half weight to the end values [31], [32]. A span of 5 was chosen for the smoothers as it provided a good balance of smoothness and resolution. Second, WC scores were calculated by the formula described by Porges *et al*. [22] and Henning *et al*. [20] in certain frequency ranges. WC of EDA was calculated under the frequency range of about 0.01 to 1 Hz. WC of IBI was calculated in two different frequency bands: low frequency (LF) band (0.04–0.15 Hz) and HF band (0.15–0.4 Hz). The HF component of heart rate variability (HRV) is also known as respiratory sinus arrhythmia which corresponds to respiratory frequency, and to which parasympathetic nerve system (PNS) activity is a major contributor [33], [34]. The LF component is affected by both the PNS and the sympathetic nerve system (SNS) [33] and is often used as an index of mental workload [35], [36]. Thus WC of IBI was divided into WC in low frequency (WCLF) and WC in high frequency (WCHF).

3) Subjective Ratings: Active users were asked to use the NASA Task Load Index (TLX) to evaluate their own workload during the task trial, while passive users were asked to use it to evaluate their group mates' (i.e., the active users) workload. The workload difference rating was then calculated by taking the absolute value of the difference between the active user's self-evaluation and the passive user's evaluation. Thus, the workload difference rating could be an indicator of how the passive user's evaluation matched the active user's self-evaluation.

Both the active and the passive users were asked to evaluate the technology's trustworthiness using a seven-point Likert scale for 12 items. The final trust rating was calculated by summing the points of the items (negative items were reversely coded). The resulting trust ratings ranged from 12 (least trust) to 84 (most trust). Two measures were derived: the average rating of trust in technology (averaging the active user and passive user's rating), and the difference rating of trust in technology (the absolute value of the difference between the active user's trust in technology rating and the passive user's rating). The difference rating is used as an indicator of the heterogeneity of trust in technology ratings between the active and passive users per group.

F. Data Analysis

For descriptive data analysis, the PC indicators were examined, and the subjective ratings were summarized. Correlation coefficients among the indicators were calculated to check if the indicators related to each other. PC measures during the BL recording and task trials were compared to see if there was a significant difference; in the BL recording, because the two participants were not working together, PC should be lower. In this comparison, the data were fitted to linear mixed effects (LME) models. An LME model is similar to a linear model in that it contains fixed effects with linear parameters but it also contains random effects to account for nonindependence in variance [37]. It is appropriate for analyzing the current study's data where there are two sources of nonindependence: participants worked as two-person groups and each group went through all three task/technology conditions. To test if PC was higher in the task sections than in the BL recording, the BL PC indicator values were subtracted from the task/technology conditions' PC indicator values. Then, the task trial PC indicator values were fitted with an LME model using random intercepts for groups, and only the intercept as the predictor.

TABLE III
CORRELATIONS AMONG THE PC INDICATORS FOR EDA

	SM	IDM	DA	CC
SM				
IDM	0.43 *			
DA	−0.10	0.21 *		
CC	0.00	0.00	0.06	
WC	0.05	−0.28 *	0.00	0.05

* indicates the correlation coefficient is significantly different from 0 with $p < 0.05$.
SM = signal matching, IDM = instantaneous derivative matching, DA = directional agreement, CC = cross correlation, WC = weighted coherence.

Next, the four proposed hypotheses were tested. To test the effect of task/technology conditions on the dependent variables, LME models were fitted to the data using task/technology conditions as an independent variable and groups as random intercepts. To test the relationship between PC and group performance, LME models were fitted to the data using performance as the dependent variable, PC indicator and task/technology conditions as independent variables, and groups as random intercepts. Thus, the effect of task/technology conditions was controlled statistically. A similar procedure was used to test the relationship between PC and the subjective ratings. *P*-values for the fixed effects were obtained through Type III *F*-tests. The denominator degrees of freedom in the *F*-tests were approximated using the Kenward–Roger method [38]. The degrees of freedom for *t*-tests used for pairwise comparisons were approximated using the Satterthwaite method [39]. To test the hypotheses, one-tailed tests were performed according to the stated direction of the relationship. Two-tailed tests were used for the tests that were not related to the hypotheses.

All of the data analysis was conducted using R [30] with the lme4 package [40] and the lmerTest package [41].

V. RESULTS

A. Descriptive Data Analysis

Cardiovascular activity data (IBI) from two groups were excluded in the analysis due to equipment malfunction and thus poor signal quality of that physiological measure. The final sample size of groups was 24 for EDA and 22 for IBI. Tables III and IV show the correlations among the PC indicators for EDA and IBI, respectively. For EDA, SM and IDM, IDM and DA, and IDM and WC were significantly correlated. The correlations were not significant among other PC indicators. Specifically, CC did not correlate with any other PC indictors. A different pattern was found for IBI. SM and IDM, SM and CC, and DA and CC were significantly correlated. DA, WCLF, and WCHF did not show significant correlation with each other and other PC indicators.

TABLE IV
CORRELATIONS AMONG THE PC INDICATORS FOR INTERBEAT INTERVAL

	SM	IDM	DA	WCLF
SM				
IDM	0.30 *			
DA	−0.15	−0.13		
CC	−0.29 *	−0.15	0.32 *	
WCLF	0.04	−0.02	0.02	0.08
WCHF	−0.18	−0.05	0.15	0.20

* indicates the correlation coefficient is significantly different from 0 with $p < 0.05$.
SM = signal matching, IDM = instantaneous derivative matching, DA = directional agreement, CC = cross correlation, WCLF = weighted coherence in LF, WCHF = weighted coherence in HF.

TABLE V
MEAN VALUES OF THE PC INDICATORS FOR EACH PHYSIOLOGICAL MEASURE IN THE BL RECORDING AND EACH TASK/TECHNOLOGY CONDITION

Measure	PC indicator	Conditions				t value
		Baseline	Normal	Hard	Low reliability	
EDA	SM	0.938	0.678	0.731	0.520	$t = -2.11*$
	IDM	0.052	0.051	0.051	0.051	$t = -0.02$
	DA	0.547	0.591	0.589	0.583	$t = 2.39*$
	CC	0.296	0.159	0.276	0.212	$t = -0.75$
	WC	0.242	0.286	0.263	0.293	$t = 0.86$
IBI	SM	1.214	0.958	1.023	0.922	$t = -5.14*$
	IDM	0.471	0.451	0.451	0.421	$t = -1.26$
	DA	0.507	0.502	0.506	0.508	$t = -0.37$
	CC	0.050	0.059	0.095	0.040	$t = 0.56$
	WCLF	0.176	0.179	0.186	0.162	$t = -0.02$
	WCHF	0.151	0.158	0.179	0.177	$t = 1.92$

The t values were obtained from tests that compared the average value in the task/technology conditions with the value in the BL recording. * indicates the corresponding MCMC-estimated p-value was smaller than 0.05.
SM = signal matching, IDM = instantaneous derivative matching, DA = directional agreement, CC = cross correlation, WC = weighted coherence, WCLF = weighted coherence in LF, WCHF = weighted coherence in HF.

The means of the PC indicators for EDA and IBI under the BL recording and each task/technology conditions are shown in Table V. The results of the LME model fitting were also reported in Table V. For EDA, higher PC in task trials was found for SM ($t(23) = -2.110$, $p = 0.046$, two-tailed), and DA ($t(23) = 2.388$, $p = 0.026$, two-tailed). For IBI, higher PC in task trials was found for SM ($t(21) = -5.139$, $p < 0.001$, two-tailed).

Table VI shows the descriptive statistics of the subjective ratings, including the workload difference rating, trust in technology average rating, and trust in technology difference rating.

The effect of gender composition of the groups on physiology compliance was tested, and the result was nonsignificant.

B. Effect of Task/Technology Conditions

The task/technology conditions had significant effects on performance in the tracking task ($F(2, 46) = 60.132$, $p < 0.001$) and the resource management task ($F(2, 46) = 142.320$, $p < 0.001$). Specifically, average performance in the tracking task was significantly lower in the hard condition than in the normal condition ($t(46) = -9.020$, $p < 0.001$, two-tailed); average performance in the resource management task was significantly lower in the low reliability condition than in the normal condition ($t(46) = -15.210$, $p < 0.001$, two-tailed). No significant effect was found for the task/technology conditions on the performance in the monitoring task. These results are visualized in Fig. 2.

TABLE VI
RANGE AND MEAN (±SD) OF THE SUBJECTIVE RATINGS UNDER EACH TASK/TECHNOLOGY CONDITION

Measurement	Range	Conditions		
		Normal	Hard	Low reliability
Workload difference rating	0.17–48.5	16.17 (±11.34)	11.12 (±8.07)	17.61 (±12.76)
Trust in technology average rating	36–65	52.21 (±7.08)	48.58 (±7.18)	51.06 (±5.52)
Trust in technology difference rating	0–25	11.42 (±6.86)	11.04 (±5.95)	10.50 (±6.56)

WCHF of IBI was the only PC indicator that was significantly affected by the task/technology condition ($F(2, 42) = 3.256$, $p = 0.0484$). Both the hard condition ($t(42) = 2.080$, $p = 0.022$, one-tailed) and low reliability condition ($t(42) = 2.320$, $p = 0.012$, one-tailed) had higher average values in WCHF of IBI than in the normal condition.

The task/technology conditions had a significant effect on the average trust in technology rating ($F(2, 46) = 7.493$, $p = 0.002$). Specifically, ratings in the low reliability condition were lower than in the normal condition ($t(46) = -3.790, p < 0.001$, two-tailed). A significant effect was also found for the workload difference rating ($F(2, 46) = 4.714$, $p = 0.014$); difference ratings in the hard condition were lower than in the normal condition ($t(46) = 2.440$, $p = 0.019$, two-tailed). These results are visualized in Fig. 3.

C. Relationship Between Physiological Compliance and Group Performance

DA of EDA was related to performance in the tracking task ($F(1, 58) = 9.384, p = 0.003$). Higher PC was related to higher performance ($t(52) = 2.541$, $p = 0.007$, one-tailed). In addition, SM of EDA was related to performance in the resource management task ($F(1, 65) = 5.871$, $p = 0.018$). However, in contrast with the previous result, lower SM of EDA (thus higher PC) related to lower performance ($t(52) = -2.160, p = 0.982$, one-tailed). These results are visualized in Fig. 4.

D. Relationship Between Physiological Compliance and the Subjective Ratings

The results indicated that none of the PC indicators correlated with the workload difference rating after controlling for the effect of task/technology conditions.

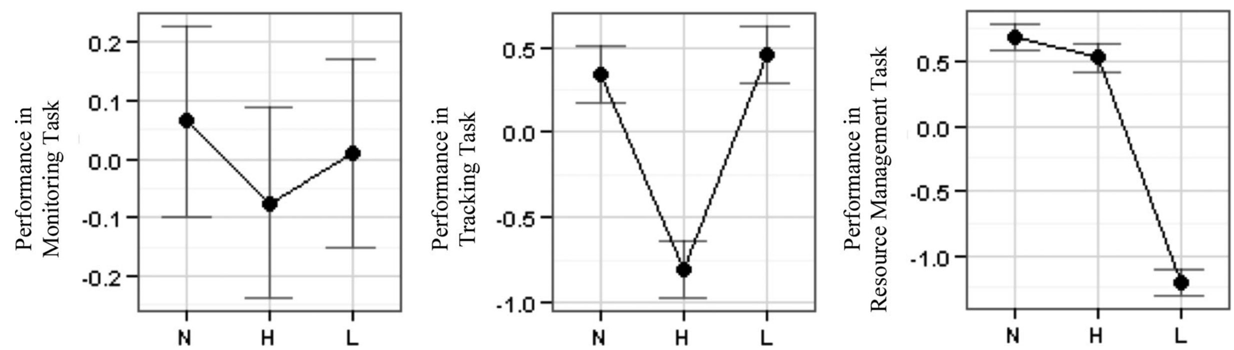

Fig. 2. Mean and standard deviation of the performance in monitoring task, tracking task, and resource management task in the normal condition (N), hard condition (H), and low reliability condition (L).

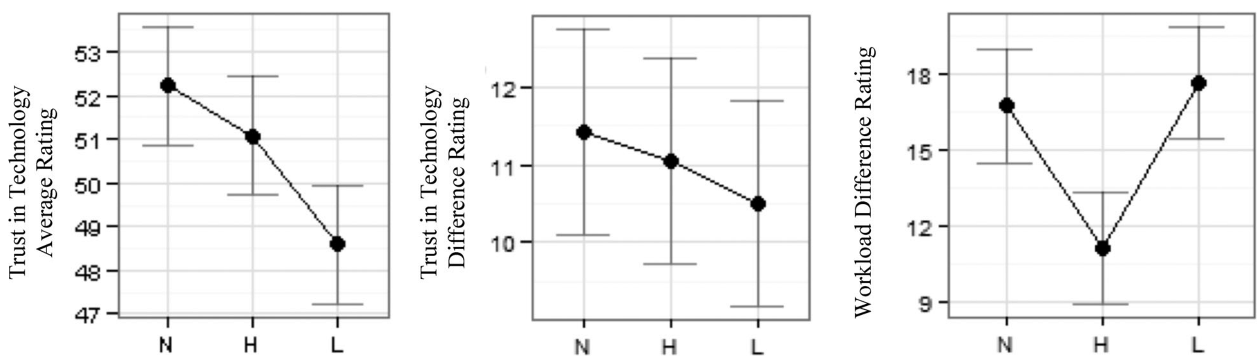

Fig. 3. Mean and standard deviation of subjective ratings, including trust in technology average rating, trust in technology difference rating, and workload difference rating under the normal condition (N), hard condition (H), and low reliability condition (L).

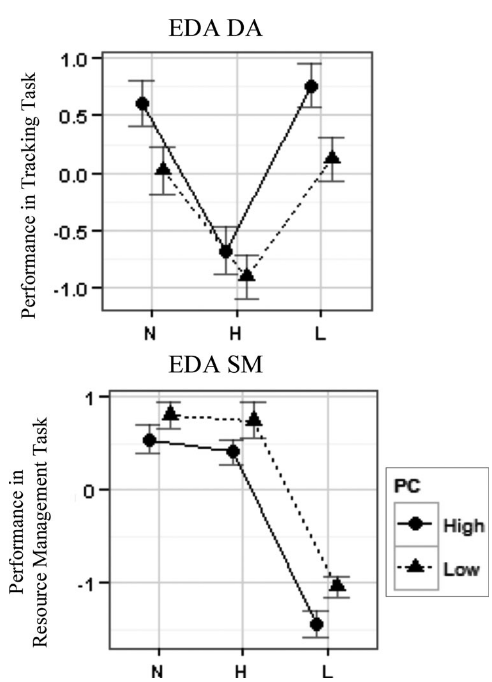

Fig. 4. Mean and standard deviation of the performance in different tasks as predicted by different levels of PC in the normal condition (N), hard condition (H), and low reliability condition (L). $+/-1$ SD value of the sample were used as high/low level of PC for data visualization purpose.

The results also indicated that PC indicators were related to the trust in technology difference rating. Levels of two PC indicators, including DA of EDA and WCHF of IBI, were correlated with values of the trust in technology difference rating (for DA of EDA, $F(1, 55) = 4.175$, $p = 0.045$; for WCHF of IBI, $F(1, 55) = 9.458$, $p = 0.003$). Higher levels of PC per the two indicators were related to lower values of the trust in technology difference rating (for DA of EDA, $t(54) = 2.110$, $p = 0.020$, one-tailed; for WCHF of IBI, $t(54) = 1.773$, $p = 0.041$, one-tailed). These results are visualized in Fig. 5.

VI. Discussion

In general, many of the different PC indicators were not significantly correlated with each other for both EDA and cardiovascular activity (IBI), which indicates a low convergent validity [42] for these measures. This may also indicate a need for developing better PC indicators for EDA and IBI in future research. However, the PC indicators used in this study reflect some unique aspects of PC. Since SM measures the area of overlap of two physiological signals, it takes the general similarity of the two signals into account with low specificity in the direction and magnitude of momentary fluctuations. Although Elkins *et al*. [21] found that SM did not show high correlation with the visual inspections of human observers of two signals,

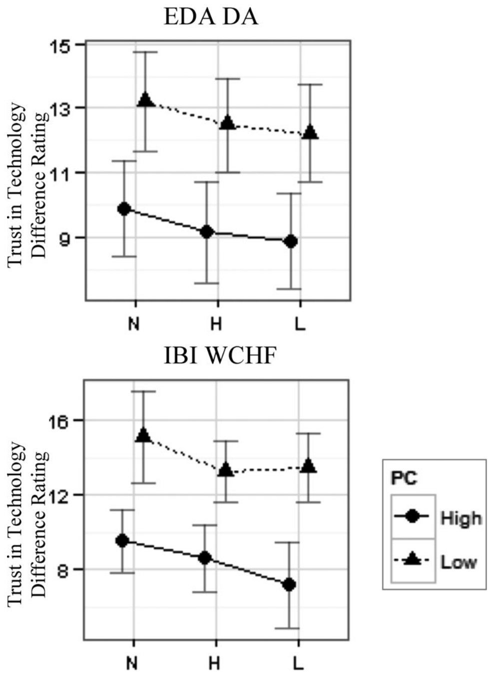

Fig. 5. Mean and standard deviation of the trust in technology difference rating as predicted by different levels of PC in the normal condition (N), hard condition (H), and low reliability condition (L). $+/-1$ SD values of the sample were used as high/low levels of PC for data visualization purposes.

this study indicated that SM successfully distinguished PC in the BL recording from the task trials. CC also measures general similarity of two physiological signals. However, it is not known if lag 0 CC is the optimum measure since there are individual differences in delays of physiological changes [43]. DA and IDM seem more sensitive to momentary changes of the signals. DA is sensitive to momentary direction changes of two signals but ignores the amplitude of the change. IDM takes both the direction and the amplitude of the change into account. These two measures may be stricter than SM and CC in complex environments and activities where compliance may not occur momentarily. In contrast with the aforementioned four time-domain PC indicators, WC is a frequency-domain measure that captures the similarity of two signals in terms of magnitudes of waves in specified frequency bands. It is useful in that it could provide information about PC in different frequency bands in HRV. However, this measure is not valid if the signals are not stationary [44]; nonstationary could be the case for EDA signals as discrete acyclic events elicit SCR in EDA.

Although there was no specific hypothesis for comparing PC during the task trials with the BL recording, it is interesting to find that only a few PC indicators show a significant difference between these two periods. During the BL recording, both of the participants in a group were instructed to sit still and relax; this may cause a relative "smooth" physiological change from both of the participants, and thus an indication of high PC.

Hypothesis 1 stated that task/technology conditions will have an effect on PC, and that high task difficulty and low technology reliability will relate to high PC. The results indicated that only WCHF in IBI was affected by the task/technology conditions. According to Boucsein and Backs' three arousal model [35], HRV is an indicator of the effort arousal system, which is responsible for inhibiting immediate response behavior to stimuli and allows for central processing of information [43], [45]. Thus, it is related to operator effort and workload [46], [47]. WCLF and WCHF of IBI could be an indicator of a shared workload level of the two individuals in a group. Other PC indicators with EDA and IBI measures are all related to the affect arousal system, which is responsible for focusing attention and generating orienting responses [35], [43], [45]. Therefore, the results of this study indicated that under varied task/technology conditions, group members had similar PC levels on indicators related to attentional and operational control, but different PC levels on indicators related to workload. Specifically, under high demand induced by either the difficulty of the task or the reliability of the technology, participants had a stronger linkage to the workload to cope with the situation.

Hypothesis 2 stated that PC is positively related to task performance after controlling for the effect of task/technology conditions. This hypothesis was not supported. Many of the PC indicators were not significantly correlated with performance, but DA of EDA was significantly correlated with performance on the tracking task. High performance on the tracking task was related to high PC, even though the passive user was not in control of the computer. One explanation for this is that the tracking task requires intense attention and operation. Since the passive user was not in control, he or she may have had more attentional resources available for monitoring abnormalities in the tracking task. If the passive user directs the active user's attention to the abnormality in the tracking task when needed, the group performance on this task could improve. This communication process between the active and passive users could lead to higher PC. Therefore, in the end, high PC was related to high performance on the tracking task. Interestingly, group performance on the resource management task was negatively related to level of PC as indicated by SM of EDA. In this experiment, the resource management task did not require intense attention and operation; rather, it required more strategic planning from the group. Further investigation is needed to confirm and understand why PC or general arousal levels would be negatively related to those kinds of tasks.

Hypothesis 3 states that PC is related to how the passive user estimates the active user's workload. This hypothesis was not supported by the results of this study.

Hypothesis 4 states that PC is positively related to the shared perception of technology trustworthiness. Support for this hypothesis was found in this study. Three PC indicators consistently showed that higher PC was related to smaller differences of trust in technology ratings between the group members. A previous study found that there is a social influence on trust in technology ratings when the technology is being used in a team setting [10]. This study found that social influence might be related to how the active and passive users interact with each other. Further research is needed to better understand the detailed mechanisms of trust in technology and the trust calibration process at the group level.

There are several limitations of this study. First, in such an experimental study, the members of the groups did not have long-term working relationships on the task environment. Although it has implication for short-term working relationships, such as in emergency care teams, more research is needed for longer term relationships such as clinician–patient relationships or organizational work teams. Second, although a unified performance index for the monitoring task has the benefit of simplifying the data analysis process, it might not be an accurate account of the overall performance in this task. Since the distribution of reaction time, miss rate, and false alarm rate was skewed, there may be biases when the values are standardized to form a single index. Third, a number of PC indicators were used in this study, making the number of tests conducted high and vulnerable to type I errors. Fourth, although the participants were motivated for task performance in the experiment according to the observation of the experimenters, there was no method implemented to ascertain if PC was indeed a result of collaboration but not socialization for all the groups. It was also not possible to assess the effects of potential preexisting relationships between participants on PC. Finally, a more effective BL design is needed to test if PC indicators can measure the physiological covariation during collaboration. In the current study, PC measured accidental physiological covariation of two individuals who sat quietly. During the task, PC measured coincidental physiological covariation, physiological covariation due to individual task work, and physiological covariation due to collaboration. A better BL measure might be able to account for the physiological covariation due to individual task effects; however, designing such a BL is particularly challenging since the participants were assigned to active or passive user roles. Future studies could address this issue by using a different role structure.

VII. Conclusion

This study investigated the process of joint activity using PC. Establishing common ground is critical for effective joint activity, and common ground could be established through communication, sharing displays, and monitoring each other. Understanding the processes and conditions leading to common ground can provide insights on team coordination, but it can be difficult to measure common ground dynamically in a complex task environment. Physiological measures can complement observation or subjective reports in understanding the processes and conditions leading to common ground. In this experimental study, participants performed group tasks under varied task/technology conditions in groups involving active and passive users. Results informed four main findings: First, different PC indicators measure different aspects of PC. Researchers and practitioners should be careful about selecting appropriate PC indicators. More research is needed to develop and compare different PC indicators. Second, PC is related to group performance after controlling for task/technology conditions. Specifically, DA of EDA is related to performance in the tracking task, and SM of EDA is related to performance in the resource management task. Results from the first two findings can inform future research seeking to measure the dynamic processes of common ground among team members operating in complex task environments. The third main finding was that task/technology conditions affect the level of PC among the group members. In a joint activity where high attentional resources are needed for operational control, such as in the tracking task used in this study, common ground may play an important role in performance; higher PC was related to higher group performance in the tracking task. It is still undetermined why common perceptions of workload may not be shared among active and passive users of technology. Technology could play a supportive role in this regard by supporting passive users' sensitivity to active users' workload, such as when a teammate is deciding whether or not to interrupt another teammate [48]. Finally, PC relates to a shared perception of trust in technology among the group members. Specifically, higher PC as indicated by DA of EDA and WC of HF bands of IBI related to higher similarity in trust in technology rating. These results support the notion that trust can be an important factor in identifying or establishing common ground among team members. It is possible that a person's trust in technology could influence another co-user's trust in technology, and efforts to establish common ground.

Appendix

Trust in technology rating scale [29]. The participants rated each item on a scale from 1 (disagree) to 7 (agree).

1) The system is deceptive.
2) The system behaves in an underhanded manner.
3) I am suspicious of the system's intent, action, or outputs.
4) I am wary of the system.
5) The system's actions will have a harmful or injurious outcome.
6) I am confident in the system.
7) The system provides security.
8) The system has integrity.
9) The system is dependable.
10) The system is reliable.
11) I can trust the system.
12) I am familiar with the system.

References

[1] H. H. Clark and S. E. Brennan, "Grounding in communication," in *Perspectives on Socially Shared Cognition*, L. B. Resnick, J. M. Levine, S. D. Teasley, Eds. Washington, DC, USA: APA Books, 1991.

[2] G. Klein, P. J. Feltovich, J. M. Bradshaw, and D. D. Woods, "Common ground and coordination in joint activity," in *Organizational Simulation*, W. B. Rouse and R. B. Kenneth, Eds. Hoboken, NJ, USA: Wiley, 2005, pp. 139–184.

[3] C. Gutwin and S. Greenberg, "The importance of awareness for team cognition in distributed collaboration," in *Team Cognition: Understanding the Factors that Drive Process and Performance*, E. Salas and S. M. Fiore, eds., Washington, DC: American Psychological Association, 2004, pp. 177–201.

[4] J. G. Bruhn, "Delivering health care in America: A systems approach," *Family Community Health*, vol. 22, pp. 79–80, 2000.

[5] M. Nasser, Z. Fedorowicz, T. Newton, C. Van Weel, J. J. van Binsbergen, and F. A. Van de Laar, "Patients record systems: Effects on dental practice and patient oral health outcomes," *Cochrane Database Syst. Rev.*, 2010. Available: http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD008606/abstract (Aug. 24, 2012).

[6] E. Mitchell and F. Sullivan, "A descriptive feast but an evaluative famine: Systematic review of published articles on primary care computing during 1980–97," *Brit. Med. J.*, vol. 322, pp. 279–282, 2001.

[7] A. M. Stiggelbout, T. V. d. Weijden, M. P. T. D. Wit, D. Frosch, F. Légaré, V. M. Montori, L. Trevena, and G. Elwyn, "Shared decision making: Really putting patients at the centre of healthcare," *BMJ*, vol. 344, p. e256, 2012.

[8] D. W. Bates and A. Bitton, "The future of health information technology in the patient-centered medical home," *Health Affairs*, vol. 29, pp. 614–621, 2010.

[9] M. R. Endsley and W. B. Jones, "A model of inter-and intrateam situational awareness: Implications for design, training, and measurement," in *New Trends in Cooperative Activities Understanding System Dynamics in Complex Environments*, M. McNeese, E. Salas, and M. R. Endsley, Eds. Santa Monica, CA, USA: Human Factors and Ergonomics Society, 2001, pp. 46–67.

[10] E. Montague and J. Xu, "Understanding active and passive users: The effects of an active user using normal, hard and unreliable technologies on user assessment of trust in technology and co-user," *Appl. Ergonom.*, vol. 43, pp. 702–712, 2012.

[11] O. Inbar and N. Tractinsky, "Lowering the line of visibility: Incidental users in service encounters," *Behaviour Inf. Technol.*, vol. 31, pp. 245–260, 2012.

[12] J. M. Bradshaw, P. Beautement, M. R. Breedy, L. Bunch, S. V. Drakunov, P. J. Feltovich, R. R. Hoffman, R. Jeffers, M. Johnson, S. Kulkarni, J. Lott, A. K. Raj, N. Suri, and A. Uszok, "Making agents acceptable to people," in *Intelligent Technologies for Information Analysis: Advances in Agents, Data Mining, and Statistical Learning*, N. Zhong and J. Liu, Eds. Berlin, Germany: Springer, 2003, pp. 361–400.

[13] R. A. Henning, A. G. Armstead, and J. K. Ferris, "Social psychophysiological compliance in a four-person research team," *Appl. Ergonom.*, vol. 40, pp. 1004–1010, 2009.

[14] M. Trimmel, S. Fairclough, and R. Henning, "Psychophysiology in ergonomics," *Appl. Ergonom.*, vol. 40, pp. 963–964, 2009.

[15] R. W. Levenson and J. M. Gottman, "Marital interaction: Physiological linkage and affective exchange," *J. Pers. Social Psychol.*, vol. 45, pp. 587–597, 1983.

[16] H. B. Kaplan, N. R. Burch, S. W. Bloom, and R. Edelberg, "Affective orientation and physiological activity (GSR) in small peer groups," *Psychosomatic Med.*, vol. 25, pp. 245–252, 1963.

[17] D. Cwir, P. B. Carr, G. M. Walton, and S. J. Spencer, "Your heart makes my heart move: Cues of social connectedness cause shared emotions and physiological states among strangers," *J. Exp. Social Psychol.*, vol. 47, pp. 661–664, 2011.

[18] J. B. Nowlin, C. Eisdorfer, M. D. Bogdonoff, and C. R. Nichols, "Physiologic response to active and passive participation in a two-person interaction," *Psychosomatic Med.*, vol. 30, pp. 87–94, 1968.

[19] T. J. Smith and K. U. Smith, "Feedback-control mechanisms of human behavior," in *Handbook of Human Factors, vol. 2*. New York, NY, USA: Wiley, 1987, pp. 251–293.

[20] R. A. Henning, W. Boucsein, and M. C. Gil, "Social-physiological compliance as a determinant of team performance," *Int. J. Psychophysiol.*, vol. 40, pp. 221–232, 2001.

[21] A. N. Elkins, E. R. Muth, A. W. Hoover, A. D. Walker, T. L. Carpenter, and F. S. Switzer, "Physiological compliance and team performance," *Appl. Ergonom.*, vol. 40, pp. 997–1003, 2009.

[22] S. W. Porges, R. E. Bohrer, M. N. Cheung, F. Drasgow, P. M. McCabe, and G. Keren, "New time-series statistic for detecting rhythmic co-occurrence in the frequency domain: The weighted coherence and its application to psychophysiological research," *Psychol. Bulletin*, vol. 88, pp. 580–587, 1980.

[23] J. MacMillan, M. Paley, E. Entin, and E. Entin, "Questionnaires for distributed assessment of team mutual awareness," in *Handbook of Human Factors and Ergonomic Methods*, N. A. Stanton, A. Hedge, K. Brookhuis, E. Salas, and H. W. Hendrick, Eds. Boca Raton, FL, USA: CRC Press, 2004, pp. 484–494.

[24] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors: J. Human Factors Ergonom. Soc.*, vol. 46, pp. 50–80, 2004.

[25] E. Montague and J. D. Lee, "Trust in health technologies," in *Handbook of Human Factors and Ergonomics in Health Care and Patient Safety*, P. Carayon, Ed., 2nd ed. New York, NY, USA: CRC Press, 2012, pp. 281–292.

[26] B. M. Muir, "Trust in automation—Part I: Theoretical issues in the study of trust and human intervention in automated systems," *Ergonomics*, vol. 37, pp. 1905–1922, 1994.

[27] J. R. Comstock and R. J. Arnegard, "The multi-attribute task battery for human operator workload and strategic behavior research," Nat. Aeronautics Space Admin., Hampton, VA, USA, Tech. Rep. 104174, 1992.

[28] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," *Human Mental Workload*, vol. 1, pp. 139–183, 1988.

[29] J. Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *Int. J. Cognit. Ergonom.*, vol. 4, pp. 53–71, 2000.

[30] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2013.

[31] P. Bloomfield, *Fourier Analysis of Time Series: An Introduction*, 2nd ed. New York, NY, USA: Wiley, 2000.

[32] P. S. P. Cowpertwait and A. V. Metcalfe, *Introductory Time Series With R*. New York, NY, USA: Springer, 2009.

[33] Task Force of the European Society of Cardiology the North American Society of Pacing and Electrophysiology, "Heart Rate Variability: Standards of Measurement, Physiological Interpretation, and Clinical Use," *Circulation*, vol. 93, pp. 1043–1065, Mar. 1, 1996.

[34] G. G. Berntson, J. T. Bigger, Jr., D. L. Eckberg, P. Grossman, P. G. Kaufmann, M. Malik, H. N. Nagaraja, S. W. Porges, J. P. Saul, P. H. Stone, and M. W. van der Molen, "Heart rate variability: Origins, methods, and interpretive caveats," *Psychophysiology*, vol. 34, pp. 623–48, Nov. 1997.

[35] W. Boucsein and R. W. Backs, "Engineering psychophysiology as a discipline: Historical and theoretical aspects," in *Engineering Psychophysiology: Issues and Applications*, R. W. Backs and W. Boucsein, Eds. Mahwah, NJ, USA: Lawrence Erlbaum Assoc., 2000, pp. 3–30.

[36] M. De Rivecourt, M. N. Kuperus, W. J. Post, and L. J. M. Mulder, "Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight," *Ergonomics*, vol. 51, pp. 1295–1319, 2008.

[37] B. T. West, K. B. Welch, and A. T. Galecki, *Linear Mixed Models: A Practical Guide Using Statistical Software*. Boca Raton, FL, USA: Chapman & Hall/CRC, 2007.

[38] M. G. Kenward and J. H. Roger, "Small sample inference for fixed effects from restricted maximum likelihood," *Biometrics*, vol. 53, pp. 983–997, 1997.

[39] F. E. Satterthwaite, "An approximate distribution of estimates of variance components," *Biometrics Bull.*, vol. 2, pp. 110–114, 1946.

[40] D. M. Bates, M. Maechler, and B. Bolker, *Lme4: Linear Mixed-Effects Models Using S4 Classes*, 2013. Available: http://cran.r-project.org/web/packages/lme4/index.html

[41] A. Kuznetsova, P. B. Brockhoff, and R. Christensen, *Lmertest: Tests for Random and Fixed Effects for Linear Mixed Effect Models (Lmer Objects of Lme4 Package)*, 2012. Available: http://cran.r-project.org/web/packages/lmerTest/index.html

[42] W. M. K. Trochim and J. P. Donnelly, *The Research Methods Knowledge Base*, 3rd ed. Mason, OH, USA: Atomic Dog, 2008.

[43] W. Boucsein, *Electrodermal Activity*, 2nd ed. New York, NY, USA: Springer, 2012.

[44] S. W. Smith. (1997). *The Scientist and Engineer's Guide to Digital Signal Processing*. [Online]. Available: http://www.dspguide.com

[45] R. W. Backs and W. Boucsein, "Psychophysiology in digital human modeling," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed. Boca Raton, FL, USA: CRC Press, 2009.

[46] J. M. Aasman, G., Mulder, and J. M. Lambertus, "Operator effort and the measurement of heart-rate variability," *Human Factors, J. Human Factors Ergonom. Soc.*, vol. 29, pp. 161–170, 1987.

[47] A. H. Roscoe, "Heart rate as a psychophysiological measure for in-flight workload assessment," *Ergonomics*, vol. 36, pp. 1055–1062, 1993.

[48] S. Hudson, J. Fogarty, C. Atkeson, D. Avrahami, J. Forlizzi, S. Kiesler, J. Lee, and J. Yang, "Predicting human interruptibility with sensors: A wizard of Oz feasibility study," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2003, pp. 257–264.

Enid Montague received the M.S. and Ph.D. degrees in industrial and systems engineering from Virginia Polytechnic Institute and State University, Blacksburg, VA, USA, in 2008, specializing in human factors and ergonomics engineering, the future professoriate, women's studies, and human–computer interaction.

She is currently an Assistant Professor with the Division of General Internal Medicine and Geriatrics, Northwestern University, Evanston, IL, USA, and the Director of the Wellness and Health Enhancement Engineering Laboratory.

Jie Xu received the B.E. degree in mechanical engineering from Donghua University, Shanghai, China, the B.S. degree in applied psychology from East China Normal University, Shanghai, China, in 2009, and the M.S. degree in industrial and systems engineering with a focus in human factors and ergonomics from the University of Wisconsin-Madison, Madison, WI, USA, in 2011. He is currently working toward the Ph.D. degree with the Department of Industrial and Systems Engineering, University of Wisconsin-Madison.

He is also a Research Assistant with the Department of Industrial and Systems Engineering, University of Wisconsin-Madison. His current research interests include group use of technology, psychophysiology, emotion and trust in technology, modeling human behavior, and interface design.

Erin Chiou received the B.S. degree in psychology and philosophy from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 2008. She is currently working toward the Ph.D. degree with the Department of Industrial and Systems Engineering, University of Wisconsin-Madison, Madison, WI, USA.

She is also a Research Assistant with the Department of Industrial and Systems Engineering, University of Wisconsin-Madison.