Int. J. Human-Computer Studies 91 (2016) 1–12

Contents lists available at ScienceDirect

Int. J. Human-Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Effects of different real-time feedback types on human performance in high-demanding work conditions ☆

Iris Cohen [a,b,*], Willem-Paul Brinkman [b], Mark A. Neerincx [a,b]

[a] TNO, P.O. Box 23, 3769 ZG Soesterberg, The Netherlands
[b] Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands

ARTICLE INFO

Article history:
Received 3 June 2015
Received in revised form
13 March 2016
Accepted 15 March 2016
Communicated by Eric Ragan
Available online 25 March 2016

Keywords:
Stress
Virtual training
Decision tools
Cognitive errors
Human task performance

ABSTRACT

Experiencing stress during training is a way to prepare professionals for real-life crises. With the help of feedback tools, professionals can train to recognize and overcome negative effects of stress on task performances. This paper reports two studies that empirically examined the effect of such a feedback system. The system, based on the COgnitive Performance and Error (COPE) model, provides its users with physiological, predicted performance and predicted error-chance feedback. The first experiment focussed on creating stressful scenarios and establishing the parameters for the predictive models for the feedback system. Participants ($n=9$) performed fire-extinguishing tasks on a virtual ship. By altering time pressure, information uncertainty and consequences of performance, stress was induced. COPE variables were measured and models were established that predicted performance and the chances on specific errors. In the second experiment a new group of participants ($n=29$) carried out the same tasks while receiving eight different combinations of the three feedback types in a counterbalanced order. Performance scores improved when feedback was provided during the task. The number of errors made did not decrease. The usability score for the system with physiological feedback was significantly higher than a system without physiological feedback, unless combined with error feedback.

This paper shows effects of feedback on performances and usability. To improve the effectiveness of the feedback system it is suggested to provide more in-depth tutorial sessions. Design changes are recommended that would make the feedback system more effective in improving performances.

1. Introduction

For professionals to be ready for work under stressful circumstances, such as crises, combat or disaster scenarios, they need proper training. An effective training method for decision making under stress is learning by experience (Andresen et al., 2001; Beach and Lipshitz, 1993; Cesta et al., 2014). Scenario-based training environments can be created in Virtual Reality (VR) which provide the realistic stressful situations (Peeters et al., 2014). VR seems to be able to elicit physiological stress responses in individuals (Busscher et al., 2011; Hartanto et al., 2014) but leaves out the risk of real-life crisis and disasters (Kinateder et al., 2014). Experiencing stress in VR enhances professionals' performances in real stressful situations (McClernon et al., 2010). Adding instructions to such training would provide more advantages, especially for the training of novices (Kirschner et al., 2006). Hence, next to the VR training, other training tools are needed to help the trainee learn to perform tasks under stress.

This paper focuses on real-time feedback that can be used during simulation-based training to learn to cope with stressful situations. The motivation behind this work lies in the benefits computers could bring in the acquisition of knowledge about the cognitive and affective processes and their outcomes in simulated stressful situations. Assisting trainees in VR could be done by incorporating decision support systems into the learning environment. Cognitive prostheses are systems that replace cognitive decision-making processes for the users (Wickens et al., 2004). They work well in a clearly defined decision making situation, but they do not seem to work successfully in uncertain situations since they can only make decisions on pre-programmed situations (Reason, 1987). Human decision processes are often ahead of the system (Cohen, 1993). People are also reluctant to being subordinate to a system (Gordon, 1988; Kontogiannis and Kossiavelou, 1999; Wickens et al., 2004). Cognitive prostheses also change the nature of a task from a decision task to learning to understand the system. Cognitive tools, however, are designed to provide support to the decision makers instead of replacing them (Wickens et al., 2004). They might be more appropriate for use in training settings as they support the user in learning a skill. In real-life settings, a cognitive tool can still help professionals to be more aware of negative effects of stress. Another reason to prefer cognitive tools over cognitive

☆This paper has been recommended for acceptance by Eric Ragan.

* Corresponding author at: Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands.
E-mail address: iris.cohen031@gmail.com (I. Cohen).

http://dx.doi.org/10.1016/j.ijhcs.2016.03.007

prostheses in uncertain environments is that they fit the decision process of professionals. Professionals do not seem to pick a decision after considering several alternatives but they rely on their experiences (Klein et al., 1986). An effective support tool should be harmonized to the decision-making processes of its users.

Effective support tools for uncertain situations should focus on skill or knowledge enhancement, or control. Biofeedback methods, for example, have been used for personal stress control by providing individuals with an insight into their physiological reactions to stress. Trainees try to reduce these reactions and over time they learn to control the physiological reactions to stress. Being more aware of one's emotional state is said to leave more cognitive resources for the task (Driskell and Johnston, 2006; Gohm et al., 2001). Some studies demonstrated biofeedback's ability to reduce stress, and consequently improve performance (Bouchard et al., 2012; Prinsloo et al., 2013), but these findings may be biased due to un-blind trials (Raaijmakers et al., 2013).

Support tools focussing on skill or knowledge enhancement generally offer three types of feedback: (1) outcome feedback states the current performance of a task, (2) cognitive feedback explains how to perform the task, and (3) feed-forward helps the user to anticipate on different decision options. When outcome performance is provided on its own, it does not seem to be effective in increasing task performance (Gonzalez, 2005; Lerch and Harter, 2001). It might still put the trainee in an unguided learning situation. Combining it with feed-forward feedback on the other hand, did result in increased task performance (Lerch and Harter, 2001). It seems important to support trainees in understanding feedback that shows performance levels. For example, confronting trainees with their error tendencies helps them to avoid making these errors (Dörner and Schaub, 1994). In situations with varying situational dynamics and teamwork interdependencies, the chances to make errors are relatively high. In such situations, errors often appear as a result of a lack of communication or inappropriate task allocation (Sasou and Reason, 1999). To address this, Kontogiannis and Kossiavelou (1999) have made a number of suggestions for making decision support tools more efficient for team decision making. For example, tools should provide information on team-strategy changes fitting to the situation. Furthermore, tools should also provide insight into event escalations and indicate when changes in communication are needed. And finally, they suggest that these tools should indicate when adaptations are needed in the task allocations and structures of team members.

The effectiveness of feedback also depends on the timing between task and feedback. First of all, mood or state-dependent learning shows that retrieval works better when people are in the same mood as they were in when they were learning (Kenealy, 1997). If feedback is delayed to after the task it is likely that a person is in another mood, i.e. no longer stressed. Secondly, trainees will interpret the feedback in the context in which it is given. As a training scenario unfolds and the context changes the interpretation of the feedback may be altered. Effective feedback is therefore offered quickly after the task but not during the performance of the skill (Anderson et al., 1995; Wickens et al., 2004). Trainees will otherwise ignore the feedback resulting in no effect or they will be distracted from the task they are performing which might result in decreased performance. Shute (2008) drew the same conclusion in her review on the length and complexity of feedback. When feedback is too long or too complex, trainees will not pay attention to it. Contrary to this finding, there are also studies that did not find an effect of length and complexity of feedback.

Based on the above literature, this paper takes the stance that trainees' performances benefit from receiving immediate (real-time) feedback about their physiological stress response, predictions about their performance, and predictions about the chances that they will make specific errors. These types of feedback let users recognize their current stress state and their behavioural consequences of stress. Such a feedback system, based on the COgnitive Performance and Error (COPE) model, proved effective when provided to Naval students working in a high-end simulator (Cohen, 2015). The experiment presented here, continues this line of research by studying the effects of the different combinations of immediate feedback on task performance.

1.1. COPE model

The COgnitive Performance and Error (COPE) model (Cohen et al., 2012, 2015) shows the influences of factors in the external world, via cognitive factors, on performances. A graphical representation of the COPE model is shown in Fig. 1. The COPE model starts with stimuli from the work content. Tasks that need to be performed have certain goals and task demands. The more difficult the expected reaction to the event, the higher the task demands.

When a task is being perceived, the primary appraisal will state the severity of potential danger. The secondary appraisal will assess the situation as either a *challenge* or a *threat*. A task will be seen as a challenge when individuals believe they are able to cope with the task. When they feel they cannot cope with the task, it is seen as a threat. Individuals also rate a task on its level of *perceived task demand*. Experiencing a stressful event or task also influences the emotional state. The PAD-model by Mehrabian (1996) divides emotional state in arousal, valence and dominance. Valence

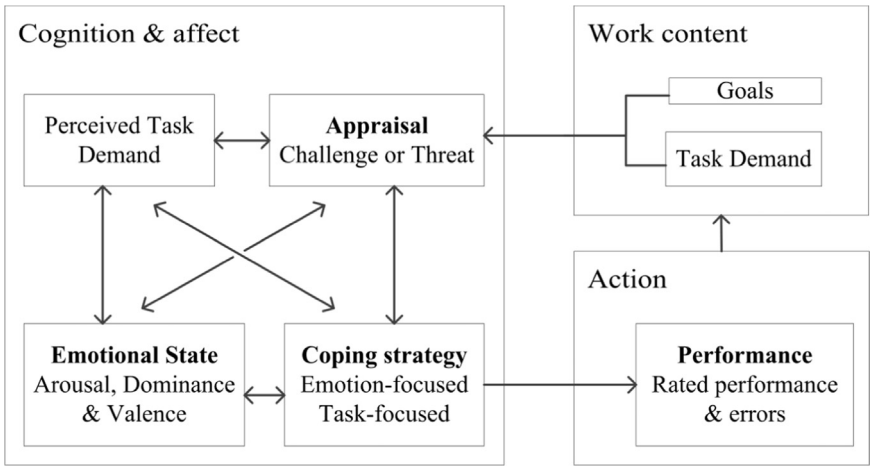

Fig. 1. Schematic view of the COPE model of external and cognitive factors, predicting an individual's performance and errors.

indicates the level of positivity or negativity, dominance indicates the level of control, and *arousal* is a level of activity. Arousal can be measured objectively with physiological factors such as *heart rate* (HR) and *heart rate variability* (HRV) (Brouwer et al., 2011; Hjortskov et al., 2004; Krantz et al., 2004).

The work content determines the appraisal, perceived task demand and arousal levels. Those variables have on their turn, influence on an individual's behaviour and actions responding to the external world. Whether a response to a task is appropriate or not, determines the level of performance. The COPE-model focusses on two types of performance: *expert rated performance*, a value given to the actions determined by experts; and the *number of errors* made during the event.

1.2. COPE feedback

Based on the COPE model, a feedback (FB) system (Fig. 2) was created that provides three types of feedback: physiological feedback (2 in Fig. 2), performance prediction feedback (1 in Fig. 2), and error-chance feedback (3 in Fig. 2). When heart rate increased, trainees could consult the other types of feedback to check what consequences the current stress had on their performance. The performance prediction feedback is a prediction about the current performance level. Although performance predictions as feedback might not be effective on its own, combined with other feedback it seems to increase performance (Gonzalez, 2005; Lerch and Harter, 2001). The third feedback type the system provided was predictions of specific error chances. The COPE model can predict four specific error tendencies (Cohen et al., 2015) namely: planning errors, communication errors, errors concerning the speed of task execution, and task allocation errors.

For all three feedback types, a higher bar graph represents a higher value. The bar graphs belonging to the same feedback type were grouped according to the principle of common region (Rock and Palmer, 1990).

The users were expected to see their heart rate increase when perceived stress increased. Instead of just focusing on reducing their physiological reactions to the perceived stress as one would do when only biofeedback is offered (Gatchel et al., 1978), they could see the consequences of the stress on their performances and adjust their behaviour accordingly.

1.3. Prototype evaluation; research questions

Although the COPE-FB system is a generic system, different scenarios require different predictive functions. Previous work showed that variables in the COPE model are influenced by tasks characteristics (Cohen et al., 2016). Therefore, the first experiment of this paper calibrates the predictive functions and also creates a set of stressful scenarios for the second experiment. In the second experiment, the COPE-FB system is used to examine the effect of providing the various types of immediate feedback by testing the following three hypotheses:

1. Immediate feedback improves trainees' performances and the perceived usability of the feedback system.
2. Immediate (a) physiological feedback, (b) predicted performance feedback, or (c) predicted error-chance feedback improves trainees' performances and the perceived usability of the feedback system.
3. Offering combinations of immediate feedback types, results in an additional positive contribution on top of the effects created by individual types of feedback, on the trainees' performances and the perceived usability of the feedback system.

2. Experiment 1: Model parametrizing

The first experiment was set out to calibrate the predictive models by determining the parameters, for the specific tasks and target groups in this paper. The first experiment also allowed to find stressful scenario's for the second study. The study was approved by the ethics committees of both TNO Soesterberg and Delft University of Technology.

2.1. Methods

2.1.1. Participants

Nine participants between 21 and 29 years old, with an average of 24 years old, participated in the experiment. Two of the participants were male. Eight of the participants were interns at TNO and all nine were students at the University of Utrecht. They were all experienced computer users and were naïve with respect to the purpose of the experiment until the debriefing.

The experimental task consisted of a simplified fire management task. It was therefore not preferred to use participants with knowledge of fire management. Instead, we recruited naïve participants who would learn to execute a stressful task, related to fire management.

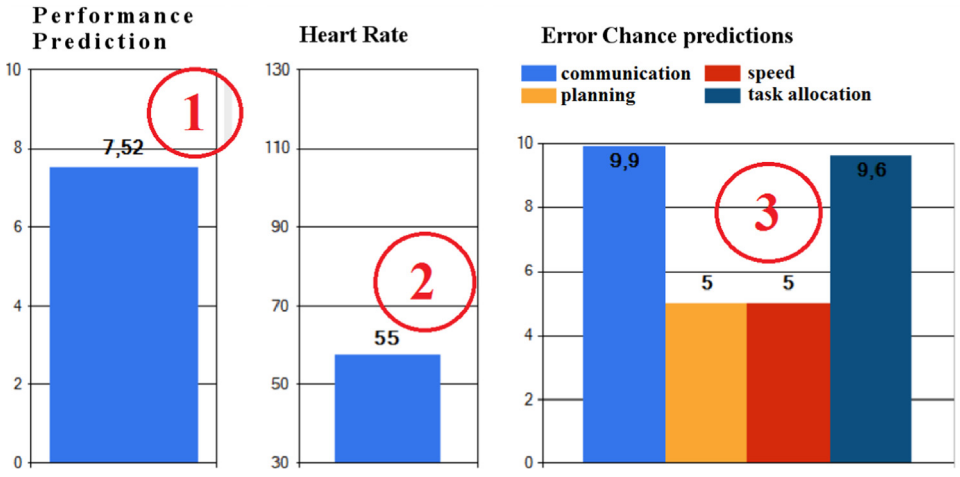

Fig. 2. Presented feedback. Performance predictions feedback on the left (1), physiological feedback in the middle (2), and predictions of error chance feedback on the right (3), with from left to right first communication error, planning error, speed error, and last task allocation error-chance feedback.

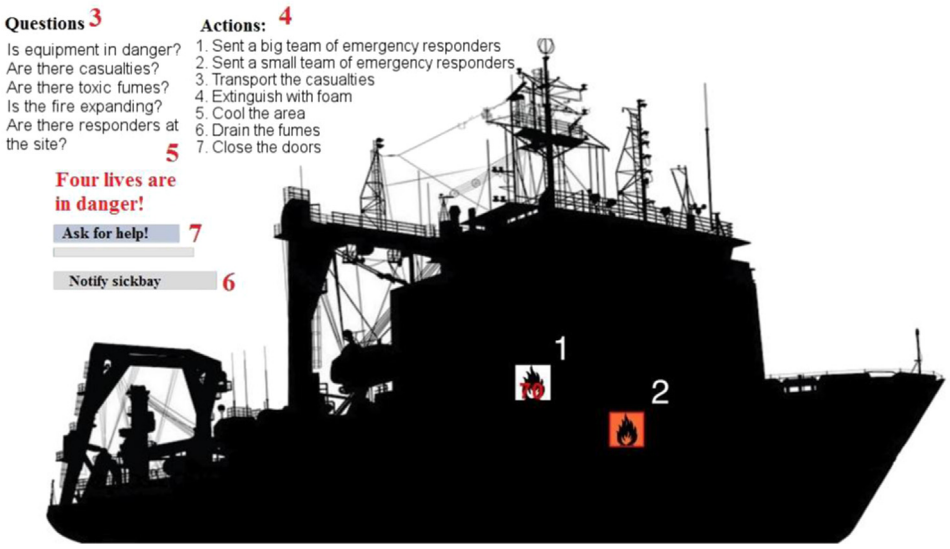

Fig. 3. Screenshot of the experimental task. When a fire occurred (1 and 2), information needed to be gathered by clicking on one of the questions (3). The answers would lead to a specific action (4) that needed to be taken in order to extinguish the fire.

2.1.2. Experimental task

Since the COPE model was validated with data collected on a Naval ship simulator (Cohen et al., 2015), a task with a similar context but lower in realism was used for this experiment. During this task, participants saw the layout of a ship on a computer screen as shown in Fig. 3. In previous experiments, this task was used to induce cognitive stress in the participants (Schreuder and Mioch, 2011). Fig. 3 shows numbers corresponding to the following aspects of the task; on the ship, two types of fires would occur: regular fires (1), represented by a white icon, as well as urgent fires (2), represented by a red icon. A normal fire had a timer that indicated how much time there was for extinguishing the fire. Urgent fires did not have a visible timer and burned down faster than normal fires, which meant that urgent fires had to be handled as quickly as possible. To create extra stress, urgent fires were accompanied by an alarm lasting 0.442 s of five equal pulses (0.05 s sound 0.05 s silence). Every pulse was a sum of sines of 520, 110 and 1458 Hz. When a normal fire had 15 s left, participants were warned by a different alarm lasting 8.19 s. This alarm consisted of 8 consecutive equal monotone pulses decreasing in amplitude and consisting of different frequencies. The signals resemble the standard British frequency Herz and pattern described by Nilsson (2014). The participants needed to collect information about the situation that could be used to determine how to extinguish the fire. This was done by selecting the fire icon followed by selecting a question (3). The answers (yes or no) would appear next to the questions.

A decision tree (Fig. 1 in the appendix) that was handed out to the participants showed the same questions. With the help of the decision tree, participants followed the questions and answers and would end at specific actions (4) that are needed to extinguish the current fire. There were four different decision trees. After four scenarios a new decision tree was used to prevent that the task could be executed automatically. The different decision trees had slightly different questions but the structure of the tree remained the same.

When a fire was selected the consequences of the fire were shown in the form of the number of (virtual) lives at stake (5). This was expected to increase the perceived stress. If a fire was extinguished, all these lives were saved. When a fire was not extinguished, all lives at stake were lost. When the workload became too high, the participant could ask for assistance (7). If the assistance option was selected for a particular fire, it disappeared from the screen and could not be used for the next 30 s. This action resulted in loss of lives to prevent participants to ask assistance for all fires. Furthermore, the assistant was not able to handle urgent fires. Some fires also required medical assistance, and the participant needed to notify the sickbay (6).

Table 1
Parameter values used to create different scenarios.

	Urgent fires		Regular fires	
	High	Low	High	Low
Time pressure	30–50 s		90–120 s	
Information uncertainty	4 s	2 s	4 s	2 s
Consequences	8 lives	4 lives	6 lives	2 lives

2.1.3. Experimental scenarios

During the experiment, scenarios were played which consisted of several fire extinguishing tasks. Scenarios were generated using a scenario generator that was given a set of three parameter values. These parameters were time pressure, information uncertainty and consequences of the decisions. The parameters could be either high or low as indicated by Table 1.

Two types of fires (regular and urgent) could occur. The consequence parameter indicated how many lives were at stake and had two values (high and low). High consequences during urgent fires meant that there were eight lives at stake and for regular fires six lives were at stake. For low consequences there were two and four lives at stake for respectively regular and urgent fires. For time pressure, there was also a distinction between regular and urgent fires. Time pressure for regular fires was 90+seconds and for urgent fires the fire needed to be extinguished within 30–50 s. Information uncertainty indicated how long it would take for information to be available. For this parameter, no distinction was made between regular and urgent fires.

The combination of all three parameter settings resulted in eight scenarios. In this experiment, participants experienced every parameter setting twice. In other words, they experienced 16 scenarios with 8 parameter settings. The order of scenarios was randomized for each participant. Each scenario lasted for about 3 min each, during which participants had to fight all fires that appeared.

2.1.4. Measurements

The COPE variables of appraisal, perceived task demand and arousal were measured for every scenario. The following subsections explain the COPE variables that were measured in this experiment to determine the predictive model parameters. To determine which scenario was stressful enough to provoke stress in the participants, perceived stress was also measured.

2.1.4.1. Appraisal (2 and 3). For measuring the appraisal that was experienced by the participants, a single 10-point scale was used. The question "I experienced the task as…" was answered from "threatening" (1) to "challenging" (10).

2.1.4.2. Perceived task demand (4). After finishing a scenario, the perceived task demand was measured using the Level-of-Information Processing (LIP) scale from the Cognitive Task Load model (Neerincx, 2003). This model consists of two other levels: Time Occupied (TO) and Task Set Switches (TSS). However, the experimental tasks were constructed in a way that the variables highly correlated, and the TO and TSS measures were incorporated in the experimental tasks (number of fires during one scenario and time pressure). In other words, the diagonal from low to high load was investigated, so that one demand indicator seemed to fulfil. The TO and TSS levels were measured as well, and used to select scenarios as described in Section 2.2.2.

2.1.4.3. Emotional state: Arousal (5 and 6). Electrocardiograph (ECG) was recorded with the Zephyr HxM. This is an unobtrusive device attached to a belt, generally used during sport. Participants placed the belt under their clothing on their chest and it sent ECG data via Bluetooth to a laptop. Heart rate (HR) in beats per minute and heart rate variability (HRV) with root-mean-square-successive-differences (RMSSD) were calculated every 10 s to assess arousal (Hjortskov et al., 2004; Krantz et al., 2004).

2.1.4.4. Performance (7 and 8). Two types of performance were measured: performance score and errors made. The performance score was related to the number of lives saved and fires extinguished. Table 2 shows the scoring method. Instead of providing participants with a number of points scored, it was called 'number of lives saved', to create a more tangible goal of the task. Another measure for performance was the number of errors made during a task. The design of the task allowed four types of errors to occur: communication errors, planning errors, speed errors and task allocation errors. For some fires, participants needed to notify the sickbay. When this action was forgotten or not performed, or when participants did not ask the right number of questions before selecting an action, a communication error occurred and one life was lost. Incorrectly asking for assistance would result in a task-allocation error. When there was an urgent fire but participants handled the regular fire first, a planning error was registered. When participants needed more than 1.25 times the average time to handle fires in a similar situation, a speed-error was registered. The average time to handle fires was calculated after the first experiment. Note that in this experiment participants did not receive immediate feedback about their errors or performance.

Table 2
Performance scoring scheme for different actions.

Action	Low consequence fire	High consequence fire
Asking for help correctly	−1	−1
Asking for help incorrectly	−All lives at stake	−All lives at stake
Notify sickbay when needed	0	0
Forget to notify sickbay	Lives saves or lost −1	Lives saved or lost −1
Extinguish a regular fire	+2	+2
Extinguish an urgent fire	+4	+6
Burn down a regular fire	−2	−3
Burn down an urgent fire	−4	−6

2.1.4.5. Perceived stress. Every scenario was rated by the participants on its stressfulness and difficulty. Perceived stress was measured with one direct question: "How stressful was this scenario". It was answered on a single 5-point scale ranging from not stressful (1), to very stressful (5). Per task, the participants filled in the Cognitive Task Load (CTL) questionnaire from Neerincx (2003). The 'Level of Information Procession', 'Time Occupied' and the 'Task Set Switches' were rated on a 5-point scale for every task.

2.1.5. Procedure

At arrival, the participants were asked to put on the heart rate monitor. The participants then read the experimental and task instructions while the experimenters checked if the heart rate monitor was working. Questions about the instructions could be asked before a tutorial trial was started with a printed version of the first decision tree. This tutorial showed the participants how to perform the task. When the task was understood, the experimental trials started. After every single scenario, participants filled in the questionnaires for the appraisal and task demand. After every four scenarios the decision tree was changed for another tree that had slightly different questions. Multiple decision trees were created to prevent participants from automatically selecting the order of the questions without reading them. A questionnaire with demographic information was filled in at the end of the experiment. This experiment lasted between 90 and 120 min.

2.2. Results

Data from this experiment was used to select the most stressful scenarios. These scenarios were used in the second experiment. The predictive functions were created based upon COPE variable data from these scenarios.

2.2.1. Data preparation

For every scenario, heart rate data, heart rate variability data and the questionnaire data were collected. Arousal data was collected every 10 s. Since the scenarios lasted approximately 3 min, this led to a list of about 18 data points per scenario. The appraisal and task demand values per scenario were the average values given by the participants in experiment 1.

2.2.2. Scenario selection

Every scenario was rated on perceived stress and the measures of CTL. The three CTL levels and the perceived stress score were used to determine the overall stressfulness of the scenarios. The median score was calculated for these variables for every scenario. Data of the scenarios with the same parameters were then combined by averaging the median scores. These average scores were used to rank the scenarios on 4 levels. Scenario 6 (*) had the highest scores on CTL, but did not score high on perceived stress. Participants had saved fewer lives in scenario 6 than in other scenarios which led to believe that scenario 6 might have been too difficult and participants gave up, which can explain feeling less stress and the decrease in performance. Therefore, scenario 6 was not selected for experiment 2. Scenarios that were selected were 8, 2, 4 and 5 since they scored high on perceived stress and on the CTL levels. These scenarios were attached to their equivalent scenario (16, 10, 12 and 15) to create new scenarios for experiment 2. Every pair created 2 new scenarios for the second experiment. For example; scenarios 8+16 and scenarios 16+8 were two new scenarios.

Table 3
Predictive performance model. The model consists of 5 variables.

Variables	Coefficient	Std. Error	t	p
Intercept	−0.028	0.081	−0.35	0.727
Arousal: HR	−0.003	0.001	−2.86	0.004
Arousal: HRV	0.013	0.003	3.95	< 0.001
Appraisal: challenge	0.033	0.009	3.54	< 0.001
Appraisal: threat	0.009	0.006	1.57	0.116
Perceived task demand	0.048	0.008	6.36	< 0.001

2.2.3. Predictive models

Due to technical problems, performance data of one participant was lost which meant that none of the data of this participant could be used for the calibration. Thus the COPE-FB System was calibrated using data of 8 participants of which 2 were male. Four Generalized Linear Mixed Models (GLMM) were created in SPSS 20.0. One model predicted performance using a linear model and three models predicted different errors using a binary logistic regression model. No planning and errors concerning the speed of the task execution were found in the dataset. Specific models could therefore, only be made for communication and task allocation errors. The fixed factors consisted of the COPE variables: HR, HRV, challenge, threat and perceived task demand. A participant-factor was included as a random factor to control for participant variation. The random effect covariance type was set to Variance Component.

2.2.3.1. Predictive Performance model. A GLMM shows that the fixed factors could explain the performance, ($F(5,24.44)=24.23$, $p<0.05$) with a weak Spearman rho correlation of $r=0.12$ between observed and predicted performance. The individual variance did not differ from the standard intercept ($var_{intercept}=0.092$, Std. Error$=0.109$, $Z=0.844$, $p=0.399$), indicating that on average the participants did not differ in their performance. Examining the coefficients in Table 3 shows that a decrease in heart rate and an increase in heart rate variability coincided with an increase in standardized performance. Additionally, an increase in challenge and perceived task demand coincided with an increase in the standardized performance.

2.2.3.2. Predictive Error models. Before predictive models could be made for the error variables, the underrepresentation of errors compared to no-errors in the dataset needed to be corrected. The error variables are binomial (0=no error and 1=error) and the observed ratio of all errors was skewed towards 0. The total error ratio was 888:30 (29.6:1), for communication errors this was 904:14 (64.57:1) and for task allocation it was 902:16 (56.38:1). By weighting the data, the ratios were stretched towards a 10:1 ratio. This ratio was chosen since it still showed a favour for 'no errors'. The total error cases were weighted with the ratio of 25:75. For the communication and task allocation errors a ratio of 15:85 was used. After applying these weightings, the new ratios were 9.87:1, 11.39:1 and 9.95:1 for the total errors, communication errors and task allocations, respectively. The predictive models were based on the weighted dataset.

The predictive model for the total error category is shown in Table 4 and is able to predict errors based on HRV, challenge and level-of-information processing $F(5, 24.44)=17.46$, $p<0.05$. A ROC curve for this model provided an Area Under the Curve (AUC) of 0.725. The individual variance did not differ from the standard intercept ($var_{intercept}=0.451$, Std. Error$=0.526$, $Z=0.858$, $p=0.391$), indicating that on average the participants did not differ in their performance.

Predictions for communication and task allocation errors can be made out of the models shown in Table 5. Communication errors could be predicted out of all the variables ($F(5, 14.744)=52.566$, $p<0.05$). A ROC curve for this model provided an AUC of 0.790. The individual variance did not differ from the standard intercept ($var_{intercept}=8.80$, Std. Error$=11.555$, $Z=0.761$, $p=0.447$), indicating that on average the participants did not differ in their performance.

Table 4
Logistic regression model that predicts the chance an error would be made. Errors are weighted with 25:75.

	Coefficient	Std. Error	t	p
Intercept	3.233	0.461	7.02	< 0.001
Arousal: HR	−0.005	0.004	−1.27	0.204
Arousal: HRV	−0.045	0.009	−5.18	< 0.001
Appraisal: challenge	0.215	0.034	6.27	< 0.001
Appraisal: threat	0.016	0.019	0.88	0.378
Perceived task demand	−0.191	0.029	−6.70	< 0.001

Table 5
Logistic regression model to prediction communication and task allocation errors[a].

Error type	Coefficient	Std. Error	t	p
Communication errors				
Intercept	9.384	1.783	5.26	< 0.001
Arousal: HR	−0.026	0.009	−3.01	0.003
Arousal: HRV	−0.093	0.011	−8.82	< 0.001
Appraisal: challenge	−0.451	0.057	−7.90	< 0.001
Appraisal: threat	−0.192	0.031	−6.11	< 0.001
Perceived task demand	−0.389	0.053	−7.35	< 0.001
Task allocation errors				
Intercept	1.626	0.912	1.78	0.075
Arousal: HR	0.000	0.004	0.08	0.938
Arousal: HRV	0.725	0.269	2.70	0.007
Appraisal: challenge	0.609	0.043	14.04	< 0.001
Appraisal: threat	0.124	0.023	5.45	< 0.001
Perceived task demand	−0.098	0.035	−2.77	0.006

[a] These errors are weighted with 15:85.

Task allocation errors could be predicted out of HRV, challenge, threat and level-of-information processing ($F(5, 14.884)=51.78$, $p<0.05$). A ROC curve for this model provided an AUC of 0.67. The individual variance again did not differ from the standard intercept ($var_{intercept}=2.474$, Std. Error$=3.367$, $Z=0.735$, $p=0.463$), indicating that on average the participants did not differ in their performance.

2.3. Discussion

The first experiment resulted in a set of 8 stressful scenarios that will be used in the next experiment. Based on data from these scenarios, four significant models were created that predicted performance, communication errors, task allocation errors and total number of errors out of the variables of the COPE-model.

The sample size of this experiment is relatively small, which limits the statistical power of the analyses. Some aspects of the parameter settings are also open for discussion. For example, the error models were based on a weighted dataset. The weightings changed the error ratios to 1:10 error, no-error ratio. Whether the ratios chosen in this experiment were satisfying depends on the interpretation of the consequences of misses and false positives in the error prediction. Since the feedback system will be used in training settings the consequences of false positives and missing errors are not severe.

Another discussion point is the 10 s interval in which the physiological measures were measured. This window was chosen because the feedback would be used in tasks that last approximately 3 min. If feedback changed every minute, only three moments of feedback will occur. Therefore, it seemed appropriate to set the predictions, and therefore the dataset, to 10 s to increase the amount of feedback moments.

For the next experiment, these models were implemented into the COPE feedback system. This system used input variables and the predictive models to calculate performance and chances on specific errors that were shown to the user this time.

3. Experiment 2 Feedback test

The second experiment focussed on the impact of different feedback types on performances. The usability of the feedback system was also investigated. The experiment was setup as a within-subjects design. Participants were provided with or without physiological feedback, performance prediction feedback, or error-chance prediction feedback. Using a full-factorial design ($2 \times 2 \times 2$), participants were exposed to eight different feedback conditions. Again, the experiment was approved by the ethics committee of TNO Soesterberg and Delft University of Technology.

3.1. Methods

3.1.1. Participants

A total of 29 participants were recruited from a participant database at TNO, a research institute in the Netherlands. People who had participated in the previous experiment were excluded. Participants were between 18 and 34 years old, with an average age of 25.5 (SD=4.67) years. Data to calculate the median age was, unfortunately, lost. Fifteen participants were male and all participants were naive with respect to the purpose of the experiment. They were compensated with 25 euros plus travel expenses. A bonus of 20 euros was awarded to the participant with the highest performance score on the experimental task.

3.1.2. Task

The same fire extinguisher task was used as in the first experiment. The participants were confronted with the eight scenarios selected in the first experiment. While the scenarios were being performed, the participants received eight different immediate feedback combinations via the COPE-feedback system.

3.1.3. Using the COPE-FB system

The models used by the COPE-FB system need five real-time input values to calculate the performance predictions and error-chance predictions. Heart rate and heart rate variability were measured real-time with the Zephyr HxM. Appraisal (challenge or threat) and task demand were rated per task in experiment 1. To use the COPE feedback system, the experimenter had to set-up the system. First, files were selected containing regression models and scenario variable values. Next, the heart rate device was connected via Bluetooth so the system could use the input signal. The experimenter then selected which parts of the feedback would be shown to the participant. When the system was running, the experimenter would select which scenario was performed. By selecting a scenario, the appraisal and task demand values for that scenario were sent from the scenario file to the regression models. As a last step, the experimenter could set the time interval for new feedback calculations. The models output was used as the feedback. The feedback was updated every 10 s.

The trainee screen only showed the output of the predictive models divided over three panels with bar graphs as shown in Fig. 2. On the left, the performance prediction was shown in one bar graph. In the middle, a bar graph showed the trainee's current heart rate. On the right, a section of predicted error chances were shown. By default, four bar graphs were shown. Above these graphs, the legend showed which graph corresponds to which error.

The first experiment in this paper explained that there was no data for planning and speed errors and therefore, no predictions about these errors could be made. The bar graphs in the feedback screen for those two errors remained therefore static on 5 (as shown in Fig. 2). If these graphs would be set to 0, participants might have thought that they were not making these errors. The participants were told that these errors would not be predicted.

3.1.4. Measurements

3.1.4.1. Performance and errors. There were two measures for performance: the total score for a task, and the number of errors made. The scoring table and error categorization were the same as in the first experiment.

3.1.4.2. Usability. After every scenario and different combination of feedback, participants were asked to judge the usability of the feedback by filling in the System Usability Scale (SUS). The SUS consisted of 10 items about the systems usability that were answered on a 5-point scale (Brooke, 1996). SUS scores have a range from 0 to 100. Next to the SUS, participants were asked to choose one of the feedback types as the most pleasant and one as the least pleasant type of feedback when the experiment was over. They were also asked to indicate why they chose those types in an open question format.

3.1.4.3. Other measures. Fig. 4 shows a participant in the experimental setting. Note that she is wearing more sensors than just the Zephyr HxM heart rate belt. For another experiment, facial movement was measured using electromyography (EMG). Data from these sensors did not enter the analyses of this paper. These sensors influenced all the participants the same manner throughout the eight conditions.

3.1.5. Procedure

At arrival, the participants put on the Zephyr HxM, read instructions and signed a consent form. The instructions consisted of an explanation of the task and the feedback system. A tutorial was started to practice the task and learn about all the options of the ship simulator. Next, the feedback screen was turned on simultaneously with the data-recording session of the COPE-FB system, and the first scenario was started. The eight different feedback conditions were counterbalanced (see table in the appendix for counterbalance order). The scenarios were all executed by the participants in the same order. After finishing each scenario, participants filled in the questionnaire about the scenario and the COPE-feedback system. After every four scenarios, the decision-tree was exchanged for another decision tree. A total of eight scenarios were performed. After the experimental task, demographic information was collected and the participants chose their most favourite and least favourite type or combination of feedback.

3.1.6. Data preparation and analyses

Heart rate scores deviating more than 2.5 SDs from the mean were considered outliers and were removed from the data file. One participant had more than 25% of the heart rate data discarded and this person was therefore excluded from all analyses. The usability analysis was therefore also based on a sample of 28 participants, of which 14 were male. For the performance analysis, the data of another three participants was discarded. During their experimental sessions technical issues with the COPE-FB system resulted in incorrect performance and error scores. This analysis was therefore based on a sample of 25 participants, of which 14 were male.

A relative performance score was used in the analyses. This score was calculated by dividing the 'lives saved in a condition', by the 'total number of lives that could have been saved in that condition'. The descriptive statistics for the performance scores and the relative performances scores are shown in Table 6. There were two types of specific errors measured during the tasks: communication and task allocation errors. The distributions of the three error variables resemble a Poisson distribution.

Fig. 4. Photograph of the experimental setup. In the foreground, a laptop shows the trainer part of COPE-FB system. In the background a participant views the ship simulation on the left display and the trainee part of COPE-FB system on the right display.

Table 6
Descriptive statistics for the (relative) performance scores.

	Performance scores	Relative performance scores
Minimum	−30	−0.94
Median	8.50	0.27
Mean	6.64	0.21
Maximum	32	1

The statistical analyses were executed in R studio. The effect of the different feedback conditions were examined using a linear mixed-effect model (LMER) function on the performance and SUS data, and a generalized linear mixed-effect model (GLMER) with the Poisson family function for error data. LMER fits linear-mixed effect models to data whereas GLMER fits generalized linear mixed-effect models to datasets. The ordinal preference data was analysed using an exact multinomial and exact binomial tests from the EMT package in R.

3.2. Results

The analysis focuses on differences in performance scores, number of errors, and perceived usability scores for all the feedback conditions. The first hypothesis states that immediate feedback in general results in an increase of performance and perceived level of usability. The second hypothesis states that the three separate feedback types increase performance and perceived level of usability. The third hypothesis states that an additional positive effect can be found on top of the effect for the separate feedback types on performance and perceived level of usability. The presentation of the results follows the order of the hypotheses.

Table 7
Likelihood ratio test for models fitting the performance scores. Testing if adding factors will improve the fit compared to the H0 model.

Model	*df*	Log likelihood ratio	χ^2	*df*	*p*
1. H0 model	3	−82.72			
2. 1+main effects	6	−82.01	1.43	3	0.699
3. 2+2way interactions	9	−81.03	3.39	6	0.759
4. 3+3way interaction	10	−79.23	6.99	7	0.431

3.2.1. Performance

Two models with performance as a dependent variable were created: a null model with no fixed factors, including only a random intercept factor for participants, and an alternative model that added a fixed two level factor (feedback, no-feedback) to the null model. A likelihood ratio test found that the model fit of the alternative-model was an improvement over the model fit of the null model ($\chi^2(1)=5.38$, $p=0.02$). Relative performance scores when no feedback was provided ($M=0.07$, $SD=0.49$) were lower than the relative performance scores when feedback ($M=0.23$, $SD=0.40$) was provided ($t(198)=2.32$, $p=0.021$).

Next, the feedback factor was split into the different feedback types. Three extra models were created that contained either (2) only the main effects of heart rate, performance predictions and error chance predictions, (3) the main effects and the 2-way interactions, and (4) the main effects, 2-way and 3-way interactions for three types of feedback. As Table 7 shows, adding the three main factors and interaction factors did not improve the model fit compared to the null model. In other words, no significant effect was found for the main effects or the interaction effects.

3.2.2. Errors

The first step of the error analysis was again to test whether feedback in general resulted in any error reduction. Again a null model and an alternative model with fixed two-level factor (feedback, no-feedback) were created. The fit of the null model did not improve when a feedback factor was added for the communication errors as shown in Table 8.

Next, the feedback factor was split into the separate feedback types and combinations as was done with the performance score analysis. Again, these models did not improve model fit compared to the null model as shows in Table 8.

3.2.3. Usability

The usability of the different feedback conditions of the COPE-feedback system were measured with the System Usability Scale and with a rating scale on which feedback was most pleasant and which one was least pleasant.

3.2.3.1. SUS scores. As with the performance and error analysis, the first step was to analyse the effect of feedback in general. Two models were again created to fit the SUS scores, a null-model and an alternative model including feedback as two-level factor. A likelihood ratio analysis found that the model fit of the alternative model was no improvement over the model fit of the null model ($\chi^2(1)=2.66$, $p=0.10$).

The second analysis step examined whether individual types of feedback or their interactions affected SUS scores. Four models were created which all significantly improved the fit compared to the null model (Table 9).

The fourth model, including main effects, 2-way interaction effects and 3-way interaction effects is analysed and presented in Table 10. The SUS score without HR feedback ($M=51$, $SD=14$) was

Table 8
Likelihood ratio test for models fitting the communication, task allocation and total error variable. Testing if adding effects would improve fit compare to the H0 model.

Error type	*df*	Log likelihood ratio	χ^2	*df*	*p*
Communication error					
1. H0 model	2	−208.33			
2. 1+main effects	5	−205.79	5.0746	3	0.1664
3. 2+2way interactions	8	−204.69	2.2051	6	0.2958
4. 3+3way interaction	9	−202.63	4.1163	7	0.1223
Task allocation error					
1. H0 model	2	−243.04			
2. 1+main effects	5	−242.86	0.3470	3	0.9510
3. 2+2way interactions	8	−241.38	3.3245	6	0.7672
4. 3+3way interaction	9	−241.27	3.5368	7	0.8313
Total error					
1. H0 model	2	−208.33			
2. 1+main effects	5	−205.79	5.0746	3	0.1664
3. 2+2way interactions	8	−204.69	7.2797	6	0.2958
4. 3+3way interaction	9	−202.63	11.396	7	0.1223

Table 9
Likelihood ratio test for models fitting the SUS variable. Testing if adding effects will improve the H0 model.

Model	*df*	Log likelihood ratio	χ^2	*df*	*p*
1. H0 model	3	−897.69			
2. 1+main effects	6	−892.90	9.58	3	0.022*
3. 2+2way interactions	9	−886.80	21.77	6	0.002**
4. 3+3way interaction	10	−886.28	22.82	7	0.002**

* $p < 0.05$.
** $p < 0.01$.

significantly lower ($t(111) = -10.77$, $p < 0.05$) then the SUS score when HR feedback was provided ($M = 55$, SD = 16). This main effect is illustrated in Fig. 5a.

Table 10 also shows a significant two-way interaction effect between HR and Error feedback. Multivariate simple effect tests were conducted to examine the interaction effect. When no error feedback was provided, SUS score was higher ($F(1, 27) = 14.43$, $p = 0.001$) in conditions with heart rate feedback ($M = 59$, $SD = 2.93$) than in conditions without this feedback ($M = 50$, $SD = 2.48$). However, when error feedback was provided, no significant ($F(1, 27) = 0.07$, $p = 0.797$) difference was found between conditions with ($M = 51$, $SD = 2.57$) or without ($M = 52$, $SD = 2.60$) heart rate feedback. A similar analysis for heart rate feedback levels was also done. When no heart rate feedback was provided, no significant ($F(1, 27) = 0.61$, $p = 0.443$) difference was found between conditions with or without error feedback. However, when heart rate feedback was provided, SUS score was lower ($F(1, 27) = 7.16$, $p = 0.013$) in conditions with error feedback than without error feedback. Therefore, as Fig. 5a shows, adding error feedback to heart rate feedback lowered the perceived usability.

3.2.3.2. Preferences rating. The participants rated which feedback type they thought was most pleasant and which the least pleasant, and gave a reasoning behind their choice. Fig. 6 shows how often a feedback condition was chosen as most or least pleasant.

With exact multinomial and exact binomial tests, it was tested if the ratings were distributed fairly over all conditions, or if they showed a preference. If there was no preference the chances for every feedback condition would be equal to 1/8.

The data showed a significant preference for one of the feedback types for both the most pleasant rating ($n = 28$, expected probability = 0.125, $p < 0.001$) and least pleasant rating ($n = 28$, expected probability = 0.125, $p = 0.003$). The condition without feedback was rated most pleasant by 41.38% of the participants. The other 58.62% choose a type of feedback as most pleasant. An exact binomial test shows that the distribution of most pleasant ratings of feedback versus no-feedback, does not show a preference ($n = 28$, expected probability 0.125:0.875, $p = 0.345$).

Table 10
Effects of feedback types on SUS scores; Model 4 including the main effects, 2-way interactions and 3-way interaction.

Effects	*df*	Sum of squares	*F*	*p*
HR	1	1017.89	7.75	0.006**
Performance	1	4.72	0.04	0.850
Error	1	307.62	2.34	0.128
HR × performance	1	26.81	0.20	0.652
HR × error	1	1265.88	9.64	0.002**
Performance × error	1	307.62	2.34	0.128
HR × performance × error	1	132.84	1.01	0.316

** $p < 0.01$.

For the least pleasant ratings, both the 'no feedback' and 'all feedback' conditions scored the highest score of 25.57%. This distribution differs from a random probability distribution ($n = 28$, expected probability 0.125:0.125:0.75, $p < 0.001$).

The participants' reasons underlying their preferences were mainly practical ones (Table 2 in the appendix). The feedback was distracting them ($n = 6$) or they did not have time to watch the feedback screen ($n = 3$). Explanations concerning the applicability of the feedback stated that participants did not understand the feedback ($n = 2$) or they thought they received too much information ($n = 2$). Surprisingly, reasons for the participants to report feedback as pleasant contradicted the reasons to dislike feedback. Participants rated feedback as useful ($n = 5$), and participants reported that they knew what to do when receiving a certain type of feedback ($n = 5$). Two participants stated that they understood the feedback and four participants even explained that they would change strategy when certain feedback was given.

3.3. Discussion

Support for the first hypothesis was only found in the performance scores. The results showed that the performance scores increased when feedback was presented to the participants. However, no support was found for this hypothesis concerning the number of errors or the SUS scores. Support for the second hypothesis was only found in the analysis of the SUS scores. The SUS score for physiological feedback was higher compared to conditions where physiological feedback was not provided. Instead of support for the third hypothesis that a combination of feedback makes a positive contribution, the findings provided grounds to, at least partly, reject this hypothesis with regard to perceived usability. Adding error-chance feedback to physiological feedback reduced the perceived usability when comparing it with a situation where only physiological feedback was provided. However, when it came to the effect on performance the findings were inconclusive on this point.

Feedback did increase the performance scores, but this effect could not be found when separate feedback types were examined. This could be due to the relative small sample size and consequently limited statistical power.

There are some limitations that should be considered when the results of this study are interpreted. One limitation concerns the explanation of the COPE feedback system to the participants. A written explanation was given to the participants. A more in-depth tutorial or a training session with the COPE feedback system might increase the participants' understanding of the system and the different feedback types.

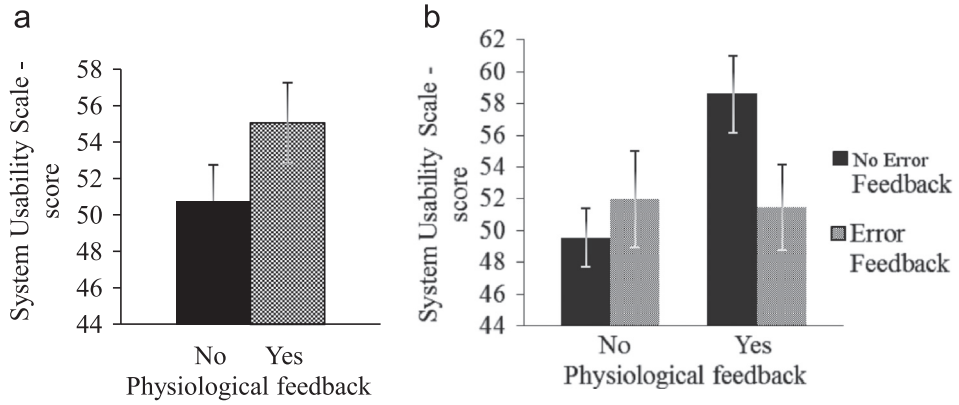

Fig. 5. (a) The main effect of physiological feedback. Average SUS score for the feedback system with and without physiological feedback with error bars showing the standard error. (b) The interaction effect of physiological × error feedback. Average SUS scores for feedback with and without physiological and error feedback with error bars showing the standard error. When error feedback is added, the main effect for physiological feedback diminishes.

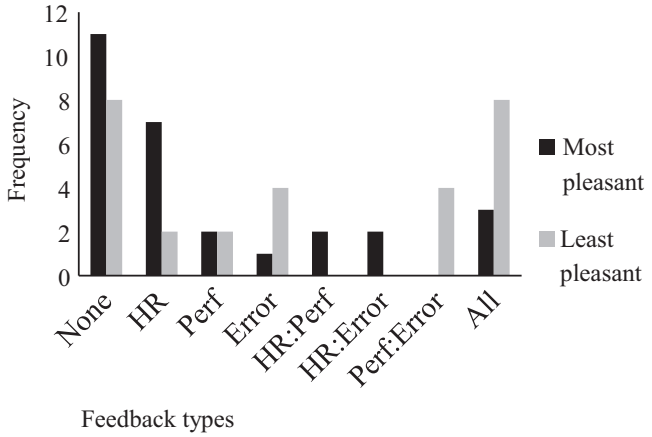

Fig. 6. Bar graphs show how often participants selected a specific feedback condition as most pleasant and as least pleasant. HR = heart rate/physiological feedback, Perf = predicted performance feedback, Error = predicted error-chance feedback.

Another limitation in the experimental design was that the COPE-FB system was designed to predict four types of errors (communication, planning, speed and task allocation errors). These errors originated from a previous study where a more naturalistic task was performed in a more complex virtual environment (Cohen et al., 2015). The computer task in this study was derived from a task that did not naturally evoke these errors. We did enrich the task in order for these errors to occur but they might have been forced upon the task. Participants in the first experiment only made two types of errors. This meant that only two of the four errors could be predicted which meant an incomplete use of the feedback system.

4. General discussion

This paper described an evaluation of the newly created COPE FeedBack system (COPE-FB system). This feedback system is based on the idea of cognitive tools providing immediate biofeedback, performance feedback and more detailed error-chance feedback (Bouchard et al., 2012; Prinsloo et al., 2013; Gonzalez, 2005; Lerch and Harter, 2001) to decrease negative effects of stress on performances. The first experiment successfully created stressful tasks and established parameters for predictive models. For the second experiment, the predictive models were implemented into the feedback system to provide participants with eight different combinations of three types of feedback. The statistical analyses showed that providing participants with immediate feedback resulted in an improvement of performance scores. However, no interaction effects of the different types of feedback were found on performances. Analysing the main and interaction effects of the different types of feedback showed an increase of System Usability Scale (SUS) score for physiological feedback over no physiological feedback. But it also showed that this improvement disappeared when adding error chance feedback to the error-physiological feedback. Overall, the usability data showed that there are good opportunities for this type of feedback to be accepted and processed for performance enhancement.

To establish such enhancement, the feedback needs substantial improvements. The current version of the COPE-FB system shows consistency in the design of the different types of feedback (Horsky et al., 2012), to rule out design effects between the feedback types. Different designs might be easier and faster to interpret. The lay-out could be designed in a way that an upward direction of the bar graphs always indicates a positive value. Currently, an increase of the error-chance bar graphs is negative, while an increase of the performance bar graph is positive. Another suggestion is to differentiate the feedback types by using different designs instead of colour schemes to make it easier for colour-blind users to differentiate between the error graphs. Participants also indicated that seeing different types of feedback can be too distracting. By implementing the system in a handheld device and, for example, adding a tactile warning signal when expected performance decreases to a certain threshold, the users do not have to constantly keep track of the feedback. In such a design using extra auditory signals should be avoided. Using alarms for both the fire-task and the feedback system could create confusion and additional cognitive load for distinguishing the source and cause of the auditory signals. The affordance (Greeno, 1994) of a tactile warning by handheld device would therefore be more effective. Another suggestion is to provide users with a more in-depth tutorial session to increase the understanding of the provided feedback. This will also increase trust in the system which is necessary in order for it to work effectively (Grootjen et al., 2006).

A limitation of this study is that the virtual task that was used lacks realism. Although the scenarios were tested on their perceived stress levels and Cognitive Task Load levels, it was not tested whether the virtual environment provoked risk perception (Kinateder et al., 2014). Although some studies question the need for training in high fidelity simulators (Beaubien and Baker, 2004;

Toups et al., 2011), others say that more realistic VR environments tend to increase the realism of the human behaviour shown in the environment (Slater et al., 2009). The results from this study might therefore not be generalizable to more realistic VR training environments. Another limitation of the design of the experimental task is that the parameter values (Table 1) and performing scores (Table 2) were arbitrarily chosen by the experimenters. The values were not tested on their realism and might therefore affect the applicability of the results to more realistic VR tasks. The lack of a realistic fire management task led to the selection of participants that had no experience in fire management. This also affects the generalizability of the results to participants that might have more experience in such tasks. Future versions of the feedback system should therefore be tested in a more realistic setting with professionals familiar in these settings, to fully investigate whether this system helps to improve performances in such situations. Furthermore, participants' individual characteristics might influence their responses to feedback. This has not been investigated in this study but might be useful to investigate in future research about this type of feedback systems.

The results of this study are promising. Performance was improved when feedback was provided and physiological feedback was preferred by the users. Further improvements might be necessary to make the COPE-FB system effective in real operational settings.

Acknowledgement

The work presented in this paper is supported by the Dutch FES programme: Brain and Cognition: Societal Innovation (project no. 056-22-010). Thanks to Maarten van der Smagt for his input as a student supervisor from University Utrecht.

Appendix

See Appendix Fig. A1, Tables A1 and A2.

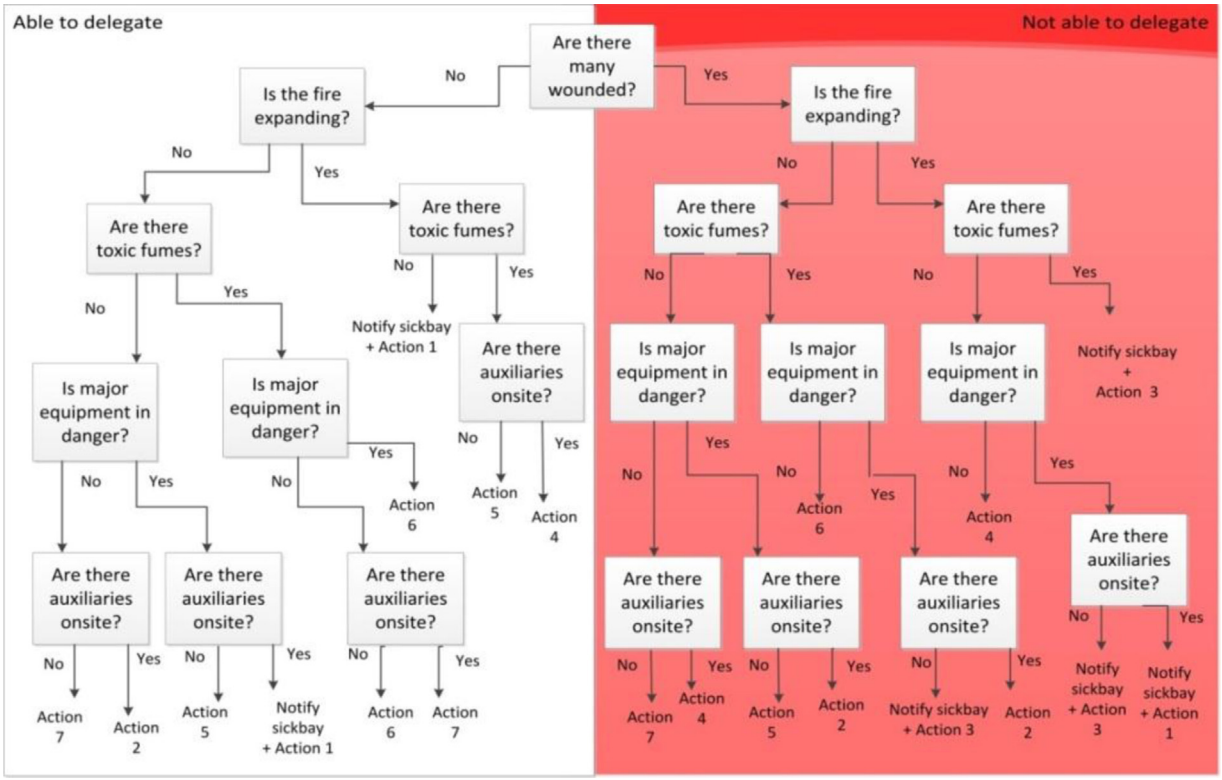

Fig. A1. Decision-tree for the fire-task. Participants had to follow the order of these questions from top to bottom. Following the correct answers results in the appropriate action to extinguish the fire. There are 4 decision-trees that were swapped after 4 scenarios. This prevented the participants from performing the task automatically.

Table A1
Orders of experimental conditions. The scenario order did not change during the experiment.

Pf:Er	Control	ER	HR:Pf	HR	HR:Er	Pf	HR:Pf:Er
Control	Pf	HR:Pf	HR	HR:Er	Pf:Er	HR:Pf:Er	Er
HR:Pf	Pf:Er	HR	Control	HR:Pf:Er	Pf	Er	HR:Er
HR:Er	HR:Pf:Er	Pf	Pf:Er	Control	Er	HR	HR:Pf
Er	HR:Er	Pf:Er	HR:Pf:Er	Pf	HR:Pf	Control	HR
Pf	HR:Pf	HR:Pf:Er	Er	Pf:Er	HR	HR:Er	Control
HR	Er	Control	HR:Er	HR:Pf	HR:Pf:Er	Pf:Er	Pf
HR:Pf:Er	HR	HR:Er	Pf	Er	Control	HR:Pf	Pf:Er

HR = heart rate feedback, Pf = performance prediction feedback, Er = error-chance prediction feedback.

Table A2
Participants' reasons to rate a feedback combination as either most or least pleasant.

Least pleasant	*n*	Most pleasant	n
No time to watch the screen	3	Useful information	5
Too much distraction	6	I understand this	2
I don't understand it	2	This is useful	2
Too much information	2	I know what to do with it	5
No added value	1	I changed strategy with this feedback	4

References

Anderson, J.R., Corbett, A.T., Koedinger, K.R., Pelletier, R., 1995. Cognitive tutors: lessons learned. J. Learn. Sci. *4* (2), 167–207.

Andresen, L., Boud, D., Cohen, R., 2001. Experience-based learning. In: Foley, G. (Ed.), Understanding Adult Education and Training. Allen & Unwin, Sydney, pp. 225–239.

Beach, L.R., Lipshitz, R., 1993. Why classical decision theory is an inappropriate standard for evaluating and aiding most human decision making. In: Klein, G.A., Orasanu, J., Calderwood, R., Zsambok, C.E. (Eds.), Decision Making in Action: Models and Methods. Ablex Publishing Corporation, Norwood, NJ, pp. 3–20.

Beaubien, J.M., Baker, D.P., 2004. The use of simulation for training teamwork skills in health care: how low can you go? Qual. Saf. Health Care *13* (Suppl. 1), i51–i56 i51-i56.

Bouchard, S., Bernier, F., Boivin, E., Morin, B., Robillard, G., 2012. Using biofeedback while immersed in a stressful videogame increases the effectiveness of stress management skills in soldiers. PLoS ONE *7* (4).

Brooke, J., 1996. SUS: a quick and dirty usability scale. In: Jordan, P.W., Thomas, B., McClelland, I.L., Weerdmeester, B. (Eds.), Usability Evaluation in Industry. CRC Press.

Brouwer, A.-M., Neerincx, M.A., Kallen, V.L., van der Leer, L., ten Brinke, M., 2011. EEG alpha asymmetry, heart rate variability and cortisol in response to Virtual Reality induced stress. J. CyberTher. Rehabil. *4* (1), 21–34.

Busscher, B., Vlieger, D. d, Ling, Y., Brinkman, W.P., 2011. Physiological measures and selfreport to evaluate neutral virtual reality worlds. J. Cyberther. Rehabil. *4* (1), 15–25.

Cesta, A., Cortellessa, G., Benedictis, R.D., 2014. Training for crisis decision making – an approach based on plan adaptation. Knowl.-based Syst. *58*, 98–112.

Cohen, I., 2015. Improving Trainees' Performances While Under Stress Using Real-time Feedback (Doctoral dissertation). Delft University of Technology, TU Delft.

Cohen, I., Brinkman, W.P., Neerincx, M.A., 2012. Assembling a synthetic emotion mediator for quick decision making during acute stress. Paper presented at the Proceedings of the 2012 European Conference on Cognitive Ergonomics, Edinburgh.

Cohen, I., Brinkman, W.P., Neerincx, M.A., 2015. Modelling environmental and cognitive factors to predict performance in a stressful training scenario on a naval ship simulator. Cogn. Technol. Work *17* (4), 503–519.

Cohen, I., den Braber, N., Smets, N.J., van Diggelen, J., Brinkman, W.P., Neerincx, M. A., 2016. Work content influences on cognitive task load, emotional state and performance during a simulated 520-days' Mars mission. Comput. Hum. Behav. *55*, 642–652.

Cohen, M.S., 1993. The bottom line: naturalistic decision aiding. In: Klein, G.A., Orasanu, J., Calderwood, R., Zsambok, C.E. (Eds.), Decision Making in Action: Models and Methods. Ablex Publishing Corporation, Norwood, NJ, pp. 3–20.

Dörner, D., Schaub, H., 1994. Errors in planning and decision-making and the nature of human information processing. Appl. Psychol.: Int. Rev. *43* (4), 433–453.

Driskell, J.E., Johnston, J.H., 2006. Stress exposure training. In: Cannon-Browers, J.A., Salas, E. (Eds.), Making Decisions Under Stress, vol. 3. American Psychological Association, Washington, DC.

Gatchel, R.J., Korman, M., Weis, C.B., Smith, D., Clarke, L., 1978. A multiple-response evaluation of EMG biofeedback performance during training and stress-induction conditions. Psychophysiology *15* (3), 253–258.

Gohm, C.L., Baumann, M.R., Sniezek, J.A., 2001. Peronality in extreme situations: thinking (or not) under acute stress. J. Res. Personal. *35*, 388–399.

Gonzalez, C., 2005. Decision support for real-time, dynamic decision-making tasks. Organ. Behav. Hum. Decis. Process. *96*, 142–154.

Gordon, S.E., 1988. Focusing on the human factor in future expert systems. Paper presented at the Artificial Intelligence and Other Innovative Computer Applications in the Nuclear Industry, Snowbird, UT.

Greeno, J.G., 1994. Gibson's affordances. Psychological Review *101* (2), 336–342.

Grootjen, M., Bierman, E.P.B., Neerincx, M.A., 2006. Optimizing cognitive task load in naval ship control centres: Design of an adaptive interface. Paper presented at the IEA: 16th World Congress on Ergonomics.

Hartanto, D., Kampmann, I.L., Morina, N., Emmelkamp, P.G., Neerincx, M.A., Brinkman, W.P., 2014. Controlling social stress in virtual reality environments. PLoS ONE *9* (3), e92804.

Hjortskov, N., Rissen, D., Blangsted, A.K., Fallentin, N., Lundberg, U., Søgaard, K., 2004. The effect of mental stress on heart rate variability and blood pressure during computer work. Eur. J. Appl. Physiol. *92*, 84–89.

Horsky, J., Schiff, G.D., Johnston, D., Mercincavage, L., Bell, D., Middleton, B., 2012. Interface design principles for usable decision support: a targeted review of best practices for clinical prescribing interventions. J. Biomed. Inform. *45* (6), 1202–1216.

Kenealy, P.M., 1997. Mood state-dependent retrieval: the effects of induced mood on memory reconsidered. Q. J. Exp. Psychol.: Sect. A *50* (2), 290–317.

Kinateder, M.T., Kuligowski, E., Reneke, P.A., Peacock, R.D., 2014. A review of risk perception in building fire evacuation. NIST Technical Note.

Kinateder, M., Ronchi, E., Nilsson, D., Kobes, M., Muller, M., Pauli, P., Muhlberger, A., 2014. Virtual reality for fire evacuation research. In: 2014 Federated Conference on Computer Science and Information Systems (FedCSIS). IEEE, pp. 313–321.

Kirschner, P.A., Sweller, J., Clark, R.E., 2006. Why minimal guidance during instruction does not work: an analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. Educ. Psychol. *41* (2), 75–86.

Klein, G.A., Calderwood, R., Clinton-Cirocco, A., 1986. Rapid decision making on the fire ground. Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Kontogiannis, T., Kossiavelou, Z., 1999. Stress and team performance: principles and challenges for intelligent decision aids. Saf. Sci. *33*, 103–128.

Krantz, G., Forsman, M., Lundberg, U., 2004. Consistency in physiological stress responses and electromyographic activity during induced stress exposure in women and men. Integr. Physiol. Behav. Sci. 2, 105–118.

Lerch, F.J., Harter, D.E., 2001. Cognitive support for real-time dynamic decision making. Inf. Syst. Res. *12* (1), 63–82.

McClernon, C.K., McCauley, M.E., O'Connor, P.E., Warm, J.S., 2010. Stress training enhances pilot performance during a stressful flying task. Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Mehrabian, A., 1996. Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament. Curr. Psychol. *14* (4), 261–292.

Neerincx, M.A., 2003. Cognitive task load design: model, methods and examples. In: Hollnagel, E. (Ed.), Handbook of Cognitive Task DesignLawrence Erlbaum Associates, Mahwah, NJ, pp. 283–305.

Nilsson, D., 2014. Design of fire alarms: Selecting appropriate sounds and messages to promote fast evacuation. Sound Saf. Soc., 33.

Peeters, M., Van Den Bosch, K., Meyer, J.-J.C., Neerincx, M.A., 2014. The design and effect of automated directions during scenario-based training. Comput. Educ. *70*, 173–183.

Prinsloo, G.E., Derman, W.E., Lambert, M.I., Rauch, H.G.L., 2013. The effect of a single session of short duration biofeedback-induced deep breathing on measures of heart rate variability during laboratory-induced cognitive stress: a pilot study. Appl. Psychophysiol. Biofeedback *38*, 81–90.

Raaijmakers, S.F., Steel, F.W., Goede, M. d., Wouwe, N. C. v., Erp, J. B. F. v., Brouwer, A.-M., 2013. Heart rate variability and skin conductance biofeedback: a triple-blind randomized controlled study. Paper presented at the Humaine Association Conference on Affective Computing and Intelligent Interaction.

Reason, J., 1987. Cognitive aids in process environments: prostheses or tools? Int. J. Man–mach. Stud. *27*, 463–470.

Rock, I., Palmer, S.E., 1990. The legacy of Gestalt psychology. Sci. Am. *263*, 84–90.

Sasou, K., Reason, J., 1999. Team errors: definition and taxonomy. Reliabil. Eng. Syst. Saf. *65*, 1–9.

Schreuder, E.J., Mioch, T., 2011. The effect of time pressure and task completion on the occurrence of cognitive lockup. In: Proceedings of the International Workshop on Human Centered Processes 2011 (HCP 2011), pp. 10–11.

Shute, V.J., 2008. Focus on formative feedback. Rev. Educ. Res. *78* (1), 153–189.

Slater, M., Khanna, P., Mortensen, J., Yu, I., 2009. Visual realism enhances realistic response in an immersive virtual environment. IEEE Comput. Graph. Appl. *29* (3), 76–84.

Toups, Z.O., Kerne, A., Hamilton, W.A., 2011. The Team Coordination Game: Zero-fidelity simulation abstracted from fire emergency response practice. ACM Trans. Computer–Hum. Interact. *18* (4), 23.

Wickens, C.D., Lee, J., Liu, Y., Becker, S.G., 2004. An Introduction to Human Factors Engineering, second ed. Pearson Education, New Jersey.