Under Pressure

Influence of different types of feedback on decision making performance in stressful situations

Sophie Zuiderduin

3490033

Applied Cognitive Psychology

30 ECTS

Thesis at TNO Soesterberg

Supervisors at TNO

drs. Iris Cohen
prof. dr. Mark Neerincx

Supervisors of the Utrecht University

dr. Maarten van der Smagt
dr. Johannes Fahrenfort

Second reader of the Utrecht University

dr. Stella Donker

July 28, 2014

Utrecht University

Contents

Abstract

People in emergency professions, such as firefighters or soldiers, often have to work and make decisions with life or death consequences in stressful conditions. To help people in emergency professions make better decisions while stressed, the 'Cognitive Feedback System' (CFSystem) has been developed. The CFSystem can support a stressed person while performing a task or making decisions by providing its users with three kinds of feedback: biofeedback, a prediction on their performance and a prediction of four types of cognitive errors stressed people have been found to make often when performing a task. The CFSystem can provide its users with more than one type of feedback at a time. The current experiment was set up to study which type of feedback or combination of types of feedback helps users improve their performance, reduces stress or is the most user-friendly. User-friendliness was measured using the System Usability Scale. Participants had to perform a fire-fighting computer simulation while they were supported by eight different combinations of feedback from the CFSystem.

Although participants did not give the feedback provided by the CFSystem high usability-scores, providing feedback did help participants improve their performance. The combination of biofeedback, performance-prediction and error-prediction was the only combination of feedback that helped participants improve their performance and was the best combination of feedback to help stressed users make better decisions. Because participants are expected to benefit more from feedback with better usability, usability of the feedback should be improved.

Results also implied that feedback was processed implicitly. Implicitly processed feedback was shown to be more helpful to participants than explicitly processed feedback.

1 Introduction

1.1 Background

Decision making has been defined as "the ability of humans and other animals to choose between competing courses of action, based on the relative value of their consequences" [1]. Under normal circumstances, a person would want to choose the course of action with the most valuable outcome. However, people do not always choose the best course of action, especially when under stress. Stress is a method of the body of reacting to a challenge. This reaction can occur when a person feels that he or she might not be able to handle a situation [18]. When people experience stress, they tend to make decisions that are less good [28].

There are many factors in decision making that can cause stress, for example when the decision has very severe consequences, when there is a high degree of uncertainty about the expected outcome [27], or when there is limited time to make the decision [19].

A higher heart rate and a change in skin conductance response are physiological consequences of stress. The psychological effects of stress depend on the appraisal a person has. When people perceive the decision or task they are facing as a challenge, they feel able to perform the task. Challenged people used their resources to focus on performing their task or to make a good decision; they are task-focused [14].

However, when people feel unable to make the right decision or unable to perform a task successfully, they feel threatened. Kassam, Koslov and Mendes [14] demonstrated that people who experience stress and feel threatened by the decision they have to make, have fewer resources to make the decision than people who do not feel threatened. Threatened people use their resources mainly to recognize and deal with the emotions they are experiencing, so they are emotion-focused. This leaves them with fewer resources for the decision.

Stress can alter the quality of a decision in many different ways. For example, stress often causes adjustments of automated responses to fail [27]. Normally when a person starts to make a decision, two processes simultaneously start to take place: The automatic process generates a quick first response. The conscious and slower decision making process can alter the automatic response, if necessary. When a person is stressed, this rational adaptation often fails. This decrease in adjustment from automated responses might be caused by fewer resources due to stress, or it might be caused by an exaggerated reliance on lower level automatic response tendencies and a decreased control over cognitive processes [27].

Furthermore, high stress-levels can cause participants to make decisions before considering all available options, which causes them to choose a suboptimal option. This was demonstrated in an experiment where participants were told they would receive an electric shock if they did not make the right decision [15], as well as in real-life situations of acute danger. Limited attention and reductions in executive functioning have been postulated to cause this so called 'dysfunctional strategy use'. [27]

Personal differences can also influence the quality of decisions a person makes when stressed. Age, gender and chronic nicotine or alcohol consumption have been shown to affect the influence of stress on decision making [27]. The ability of a person to identify and distinguish between current feelings has also been demonstrated to cause a higher decision performance, because people who can easily identify their emotions are more able to control the biases that feelings possibly cause [25].

Also, a better understanding of their own emotional state helps individuals to experience fewer cognitive difficulties when stressed. People who can easily distinguish and identify their own emotions need fewer cognitive resources to understand their current state and more resources are available for decision-making. This causes better performance under stress [9]. These people probably suffer less from emotion-focused coping strategies when confronted with a threatening task.

People in emergency professions such as firefighters or soldiers are often responsible for the lives of other people. Sometimes their own life is in danger as well. In these stressful circumstances they have to make decisions that can save or kill people. Those decisions often have to be made very quickly and with much uncertainty about the outcomes of different options.

1.2 Decision support

To help people in emergency professions make good decisions, even when they are stressed, the project 'Accommodating of emotions during quick decisions' was set up. The goal of this project is to find a way to make people in emergency professions more able to handle emergency situations . Computer programs have been suggested to assist people while handling stressful situations. Those support systems were called Intelligent Decision Aids (IDA's) [17]. The term 'IDA' denotes knowledge-based decision-support systems or software agents that can have adapting interfaces, maintain knowledge about users and generate explanations.

Some characteristics have shown to be important for IDA's to be helpful. For example, for the feedback or suggestions of the application to be useful to the user, the user has to trust the application to give accurate information [10]. Users that had no faith in the advice the application gave them or actions the application proposed, were not expected to benefit from using the application while making decisions in stressful situations. Trust grows over time when an application proves to be accurate [13]. Also, users are expected to trust the information or feedback an application provides more, if he or she can review how the feedback was established [13]. Horsky and colleagues [13] also suggested that users should be given the option to review underlying structures of the feedback instead of always showing this information to prevent an overflow of information. To prevent participants from skipping any information the application provided, it has been suggested that only relevant information is shown, and not too much at a time. [11] [13]

The importance of consistency in terminology as well as the lay-out of the IDA was emphasized by Horsky et al. [13], because in stressful situations, users should not have to use much time or cognitive resources to understand the feedback. IDA's are expected to help their users more when they are very well integrated into the task [13] . One may even speculate that if

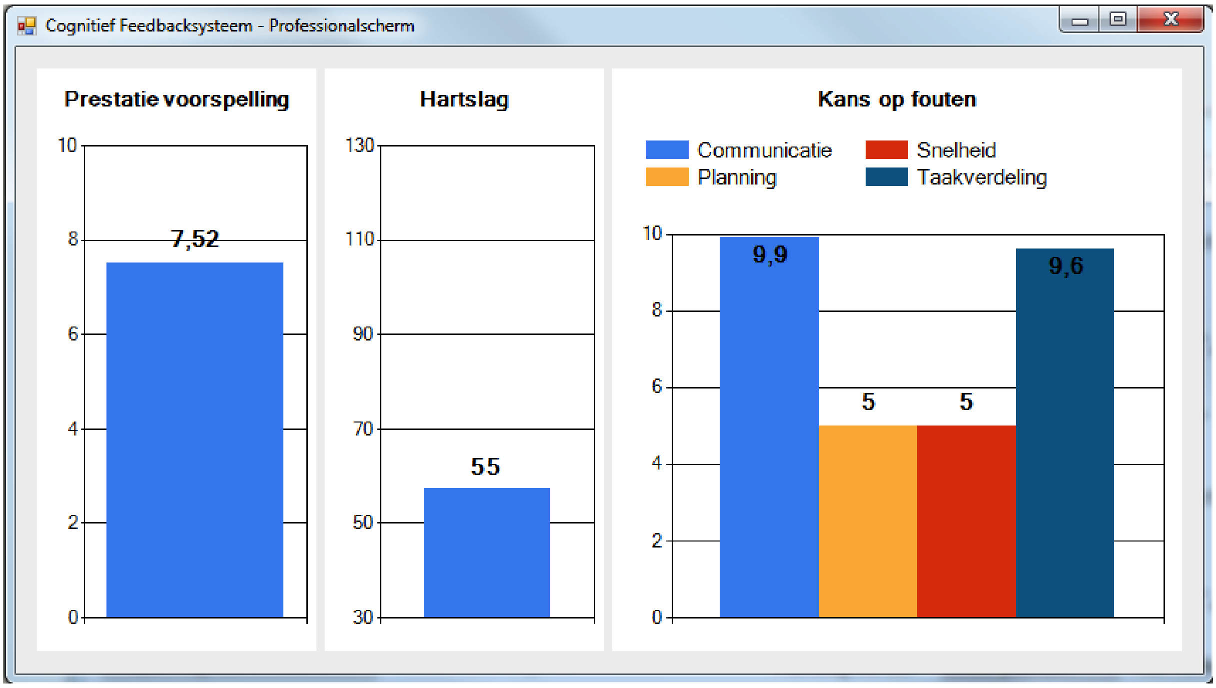


Figure 1: User interface of the Cognitive Feedback System (CFSystem) that provides users with graphs of their current heart rate, predictions of their performance and their odds of making certain errors. Translations of terms used in this Figure can be found in Appendix E.

the application scores low on usability, using the IDA can even cause more stress and a lower performance than when people perform the task without an IDA.

.

1.3 Cognitive Feedback System

The 'Cognitive Feedback System' (CFSystem) developed by Cohen, Brinkman and Neerincx [4] is an example of an advisory IDA that can provide its users with three types of feedback, while they are making difficult decisions or performing a stressful task. The CFSystem can provide participants with biofeedback, a performance-prediction and an error-prediction. Presenting stressed people with feedback, can make them aware of stress, which is expected to improve their performance [4]. The CFSystem can show participants more than a single type of feedback at a time. Figure 1 shows what the three types of feedback look like.

Showing people their heart rate is a form of biofeedback, which has been demonstrated to help participants in stressful situations by making them more aware of their physical state and thereby enabling them to control their physical stress responses [2]. One might speculate that biofeedback could also enable participants to adopt a strategy to cope with the stress, such as making decisions more carefully.

The performance-prediction showed participants how well they were predicted to perform in their current state. The performance-prediction and the error-predictions were calculated every 10 seconds, using the heart rate of a participant at the start of the 10-second interval. One might speculate that high performance-predictions reduce the stress the participant experience. However, low performance-predictions are speculated to increase stress and/or cause the participant to think longer before selecting an action.

The error-prediction made participants aware of the type of mistakes they were expected to make, which enabled them to avoid those mistakes. Communication-, planning-, speed-, and task allocation errors were predicted, because these are the four types of errors that have been shown to occur frequently when participants were stressed [4] [7]. The error-prediction provided much information to participants, so participants were speculated to make fewer mistakes, but they were also expected to make decisions slower. One might even speculate that using resources to process the feedback instead of performing the task could lead participants to pay too little attention to other aspects of the tasks and start to make mistakes other than the one they are trying to avoid.

The different types of feedback help participants in very different ways. When more than one type of feedback was presented, participants were expected to benefit all profits associated with presented types of feedback. Because showing underlying structures of predicted performance makes the predictions more reliable [11], showing biofeedback combined with performance-prediction or error-prediction was expected to make the feedback more reliable. However, more feedback was also expected to be harder to interpret and process [13], and one might speculated that the combination of all three types of feedback causes an overflow of information, which is known to reduce the ability to learn [29].

1.4 Current research

This research examines if different types of feedback can improve performance more effectively than other feedback, when working in a stressful situation. To answer this question, participants had to perform a task while receiving (combinations of) different kinds of feedback. Performance of participants with different types of feedback was compared to find out which type of feedback helped stressed participants achieve the highest performance, lowered stress or was most user-friendly.

The first expected outcome is that feedback helps participants achieve a higher performance. Furthermore, the combination of biofeedback and a performance-prediction is expected to help participants achieve the highest performance. When receiving this combination of feedback, participants were supported by the CFSystem, without receiving too little or too much information and participants were expected to trust the feedback.

If the CFSystem does improve performance of participants, emergency workers could use the CFSystem during training to learn to deal with stress better. By using the CFSystem to gain more experience dealing with emergency situations, they might be more aware of their reaction to stress. In real emergency situations, this experience could be used to deal with the situation more successfully.

2 Method

2.1 Cognitive Performance and Error model

To generate feedback that the CFSystem presents to users, Cohen and colleagues [4] designed the COgnitive Performance and Error (COPE-)model. Using this validated model it is possible to calculate how well people are predicted to perform a certain task. This prediction is calculated using physiological stress responses of a person and cognitive qualities of the task. Using heart rate as a physiological stress response and the task's appraisal and subjective task demand, the COPE-model strives to predict a performance-rating on a scale from 0 to 10 and a prediction of four types of cognitive errors stressed people have been found to make often when performing a stressful task.

The task demand of a task is determined using the Cognitive Task Load-model of Neerincx, Kennedie, Grootjen and Grootjen [22]. This model has been designed to estimate the cognitive task load of tasks, based on time occupied (TO), level of information processing (LIP) and task set switches (TSS). TO is high when maximum cognitive processing speed is required, LIP indicates whether information is processed automatically or when a task is highly cognitive demanding and TSS indicates the amount of attention shifting the task demands.

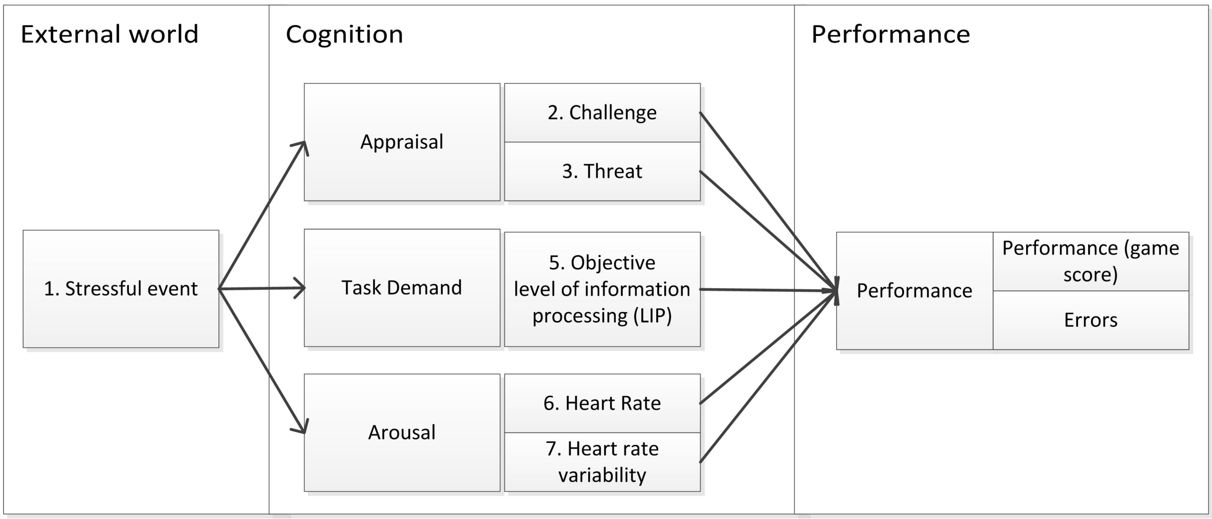


Figure 2: The COgnitive Performance and Error (COPE) model of Cohen et al. [4] calculates a prediction of performance from environmental and cognitive aspects.

Figure 2 shows the general COPE-model. This model can be calibrated by giving it all information about tasks of which it should predict the performance of participants. The model needs the name of a task, the appraisal and task demand of a task, the heart rate and heart rate variability of participants while performing the task and the accompanying performance and errors that were made. Task demand and appraisal had to be judged by people who performed the same tasks. They had to judge TO, LIP and TSS of a task, and whether they perceived the task more as a challenge or as a threat Performance accompanying a certain heart rate can be determined by ratings of professionals or by the number of points participants scored with a certain heart rate. The number of communication-, planning-, speed-, and task allocation errors participants made with accompanying heart rate and heart rate variability are necessary to predict those types of errors.

Using a calibration data set, the connections between challenge, threat, LIP, heart rate, heart rate variability and performance, and their coefficients are calculated. Connections between challenge, threat, LIP, heart rate, heart rate variability and errors, and their coefficients are also calculated. Only significant connections are used to calculate performance and error predictions. After calibration, the COPE-model needs the current heart rate of the user and the COPE-model needs to know which task is being performed and to provide feedback.

2.2 Experimental task

Initially, the COPE-model and CFSystem were designed using data collected in a naval simulator. This study therefore tries to recreate a naval setting for the stressful task. A naval fire-fighting computer simulation that Meij [21] and Schreuder and Mioch [24] used to investigate cognitive lockup has been shown to have the ability to inflict stress on participants and was therefore suitable for this experiment. The task consisted of handling fires on a ship, by finding the correct method of dealing with a fire before it got out of hand, and could be performed by participants that had no experience with fire-fighting or naval professions. Figure 3 shows how participants perceived the ship.

Fires could start at any location on the ship. The symbols next to numbers 1 and 2 in Figure 3 were used to indicate fires. Participants had to click on a fire to select it, and gain information about a fire by asking (clicking on) questions underneath number 3. After a time-interval between 0 and 8 seconds, the answer to the question, yes or no, appeared behind the question. Answers could be used to select the right action. The buttons underneath number 4

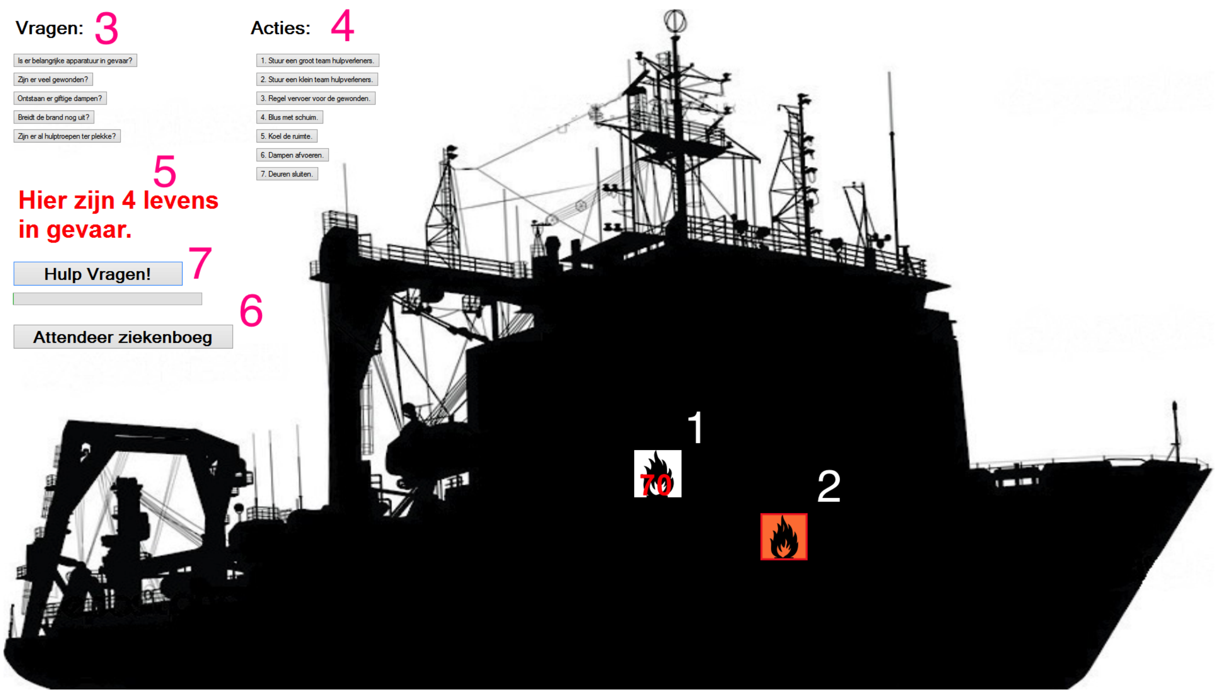


Figure 3: User interface of the firefighting task. There are currently two fires of which fire 1 is a normal fire and fire 2 is an urgent fire. Translations of important terms used in this Figure can be found in Appendix E.

could be used to select an action. Figure 4 shows a decision tree that participants used to find the correct action to handle a fire. Not all questions are translated, because their meaning is not relevant, but an example of a question is "Is it an oil fire?". If a participant had managed to select the right action in time, the fire disappeared from the screen and all lives at stake were saved. If the participant did not select the correct action before a burn down happened, all lives at stake were lost. When participants selected a wrong action, the screen froze for five seconds. This increased the time pressure and decreased the chances of the participant preventing a burn-down.

There were two types of fires: normal fires and urgent fires. Fire 1 is a normal fire and fire 2 is an urgent fire. Normal fires had a timer that indicated when they would burn down. In the situation in Figure 3, the user had 70 seconds left to handle the normal fire. Urgent fires did not have a visible timer and they burned down more quickly than normal fires, which meant that urgent fires had to be handled as fast as possible. To cause extra stress, urgent fires were accompanied by an alarm and when there were less then fifteen seconds left to handle a normal fire, participants were warned by a different alarm.

When dealing with urgent fires, participants could save or lose more lives than with normal fires. The text underneath number 5 indicates the number of lives that could be lost or saved. Table 1 gives an overview of how many lives participants could save or lose with different actions. The goal for participants was to save as many lives as possible.

The software used by Meij [21] and Schreuder and Mioch [24] could not be located, so a new program was developed for this experiment. This caused changes in lay-out and underlying structure, but not many changes in the task itself. Similarities with the task of Schreuder and Mioch [24] were the possibility for more than one fire at a time to exist and there were two types of fires: normal and urgent fires. Furthermore, the interface was changed to look more like a ship, there were different time-intervals before the answer appeared when a question was asked, alarms were added and instead of scoring points by performing the task correctly, participants had to 'save lives'. To prevent participants from performing the task automatically and without

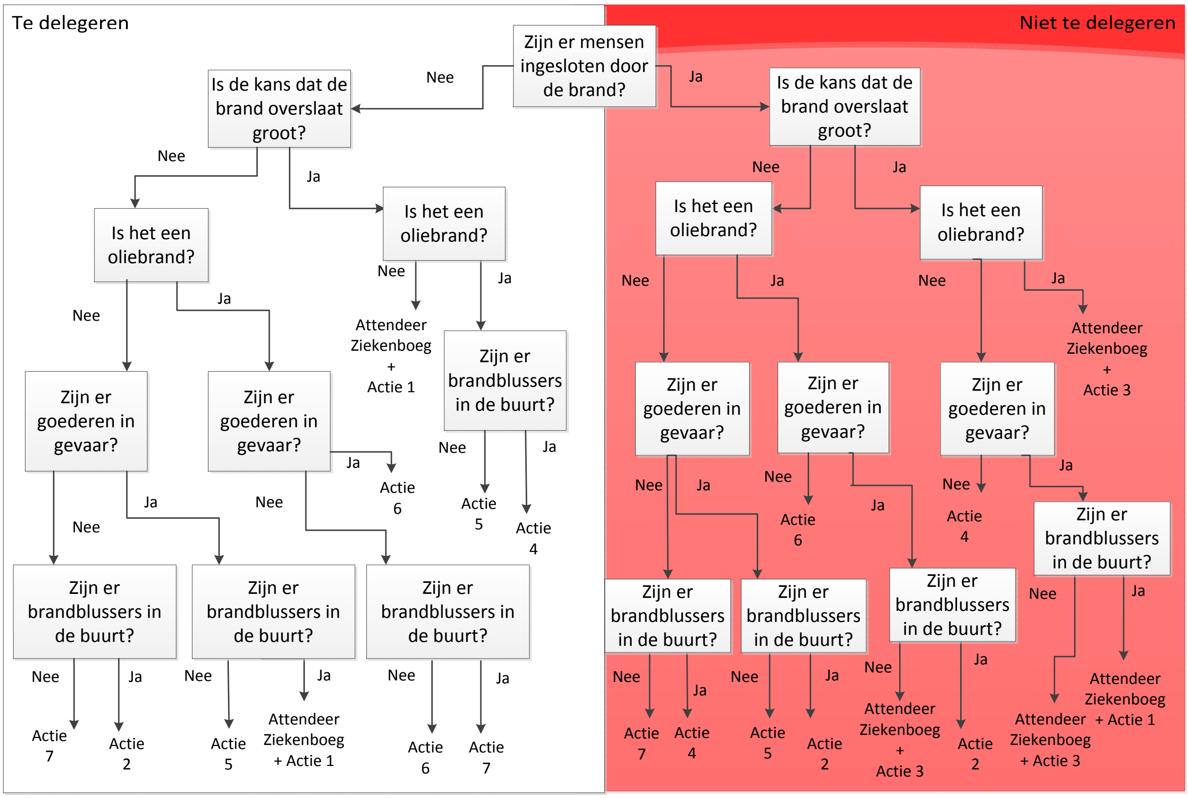


Figure 4: To find the correct action to handle a fire, participants had to ask the questions and follow the decision tree. Normal fires in the white part of the decision tree could be passed on to the 'partner'. If action 1 or 3 was the correct action, the sickbay had to be notified. Translations of important terms used in this Figure can be found in Appendix E.

Table 1: Number of lives participants could save or lose with different actions.

Action	low consequences fire	high consequences
Asking help correctly	-1	-1
Asking help incorrectly	- all lives at stake	- all lives at stake
Notify sickbay	0	0
Forget to notify sickbay	lives saved or lost -1	lives saved or lost -1
Extinguish normal fire	2	2
Extinguish urgent fire	4	6
Burn down normal fire	-2	-3
Burn down urgent fire	-4	-6

thinking, the questions in the decision tree changed every few trials. Moreover, the order in which the questions were presented was different for every fire, so participants had to keep reading the questions before clicking them, to make sure they asked the right question. This prevented the task from turning into a knowledge based task.

The task was also adapted to enable participants to make all four types of mistakes the CFSystem predicts. For some fires (those were marked on the decision tree) participants were asked to notify the sickbay. By clicking the button next to number 6, the sickbay could be warned. When participants forgot to do so or when participants did not ask the right number of questions before selecting an action, a communication error was counted and one life was lost. To enable participants to make task-allocation errors, a fictional partner was introduced. Participants that felt unable to handle all current fires, had the option of passing a fire on to the fictional partner by clicking the help-button. This is the button next to number 7 in Figure 3. To prevent participants from passing on all fires to the partner and thus reducing stress too much, passing on a task always resulted in a loss of lives, and the 'partner' was not able to handle urgent fires or fires in the red part of the decision tree. Whether the participant asked for help correctly or incorrectly, the fire disappeared from the screen and clicking the help-button was not possible for the next thirty seconds. If a participant asked for help with a normal fire in the white part of the decision tree, the number of lives lost was 1 instead of all lives that were at stake. When there was an urgent fire but participants handled the normal fire first, a planning error was registered. When participants needed more than 1,25 * the average time to handle fires in similar situations, a speed-error was registered.

2.3 Calibrating the COPE-model

The firefighting task needed scenarios that specified at what times and locations fires would start, accompanied by answers to the questions and the correct action to handle the fire.

In order to calibrate the CFSystem, to select scenarios that evoked the right amount of stress, and to be able to determine when a speed-error was made, another experiment was set up. This calibration experiment is referred to as experiment 1.

3 Experiment 1

3.1 Method

3.1.1 Ethics statement

The study was approved by the ethics committee of TNO Soesterberg. Participants were aware of the voluntarily nature of participation and they were fully debriefed about the purpose of the experiment. All participants were at least 18 years old.

3.1.2 Participants

Nine participants who were between 21 and 29 years old, with an average of 24 years old, participated in this experiment. Two of the participants were male. All participants were naive with respect to the purpose of the experiment.

Eight of the participants were interns at TNO and one of the participants was a student of the University of Utrecht. All participants were highly educated and experienced computer users.

Due to technical problems, heart rate- and performance-data of one participant were lost. The lost data could not be used to calibrate the CFSystem, thus the CFSystem was calibrated using data of 8 participants of which 2 were male.

3.1.3 Experimental design

The design of this experiment was a full-factorial within-subject design, which means that all participants were exposed to all conditions and the conditions consisted of all possible combinations of parameters.

The independent variables in this experiment were:

- time pressure
- consequences
- uncertainty

The dependent variables in this experiment were:

- appraisal (subjective variable, measured using a questionnaire)
- stress-level (subjective variable, measured using a questionnaire)
- Task load (subjective variable, measured using a questionnaire)
- performance
- heart rate
- planning errors
- communication errors
- task allocation errors

3.1.4 Set up

Apparatus and materials The experiment was conducted in a Faraday cage at TNO Soesterberg. The experimenters sat in a separate room.

The experimental task was run on a desktop computer with Windows 7. Two screens and two mice were connected to the computer. One screen and one mouse were placed on a desk in the room of the participant and one screen and one mouse were placed in the room of the experimenters. Both the experimenters and the participant were able to control the computer with a mouse, but only the experimenters had a keyboard.

The feedback participants were presented with was generated by the Cognitive Feedback System (CFSystem) of Cohen et al. [4], which was run on a Windows 7 laptop that was connected to an extra screen. Feedback was presented on the extra screen and the laptop screen which was not visible for participants was used to control the CFSystem. Heart rate was measured using a Zephyr HxM heart rate monitor that was worn around the chest by participants and connected to the CFSystem via bluetooth.

A second Windows 7 laptop was used by participants to fill in the questionnaire online. The laptop was connected to the internet via a wired connection. Headphones were worn by participants to cancel out surrounding sounds and to enable participants to hear the alarms and warning sounds from the task.

Participants were given a hard copy of the instructions, and a decision tree printed out on paper with a4-format. In Figure 4 an example of a decision tree can be found. The instructions - in Dutch - can be found in appendix A.

Table 2: Test conditions

Test condition	Time pressure	Uncertainty	Consequences	Scenarios
Low-Low-Low	Low	Low	Low	1 and 9
Low-Low-High	Low	Low	High	2 and 10
Low-High-Low	Low	High	Low	3 and 11
Low-High-High	Low	High	High	4 and 12
High-Low-Low	High	Low	Low	5 and 13
High-Low-High	High	Low	High	6 and 14
High-High-Low	High	High	Low	7 and 15
High-High-High	High	High	High	8 and 16

Scenarios To create scenarios that evoked enough stress without causing participants to just give up, the following factors that cause stress were varied between scenarios: time pressure, uncertainty and consequences. Time pressure could be high (around 30-50 seconds to handle a fire) or low (90 seconds or more to handle a fire). Uncertainty could be high (participants had to wait an average of four seconds before answers to questions appeared) or low (participants had to wait an average of two seconds before answers to questions appeared) and consequences could be low (two or four lives at stake) or high (six or eight lives at stake).

There were eight different combinations of these parameters, which were used to create eight scenarios that differed in how stressful they were for participants. To be able to calibrate the CFSystem, two sets of eight scenarios were created. These two sets were theoretically the same, because scenario one and nine, scenarios two and ten, scenarios three and eleven, etc. had the same parameters. This did not mean that these scenarios were exactly the same; for example, the location of fires and correct actions to solve fires could differ. Table 2 shows parameters of all eight scenarios.

Scenarios were generated using a scenario generator. This generator was given a set of parameters and then created eight scenarios of tree minutes, with different combinations of the parameters. Scenarios were tested multiple times before they were selected for experiment 1 and sometimes scenarios were adjusted by hand, to make them more or less stressful or to prevent all fires from having the same action as its correct action.

It was made sure that scenarios differed in their level of difficulty, and in general scenarios were challenging and stressful, but not impossible. However, one scenario that was fairly easy and one scenario in which it was impossible to save all lives were also included in the experiment.

Feedback Biofeedback was given to the participants either in the first or in the second half of the experiment, by presenting them with a graphic display of their heart rate. By having participants perform each scenario with a certain combination of parameters twice, more data was collected which enabled a better calibration of the COPE-model. Moreover, if performance changed dramatically when feedback was presented, calibration solely based data of participants who performed a task without feedback, would not give participants reliable feedback. Because half of the data was based on performance with feedback, the CFSystem would still provide reliable feedback when participants were presented with feedback.

The CFSystem was connected to a heart rate monitor and provided biofeedback as is shown in Figure 5. Participants were explained that the graph showed their current heart rate and what they could do with the information provided by the CFSystem. If participants noticed a higher heart rate, they were advised to check with themselves if they felt stressed. If that was the case, participants were advised to act more carefully, because stressed people often make more mistakes. Participants could also try to calm themselves down. The graph displayed a persons heart rate on a scale from thirty to one hundred and thirty. Participants could see their heart rate change and had to determine for themselves whether their current heart rate was

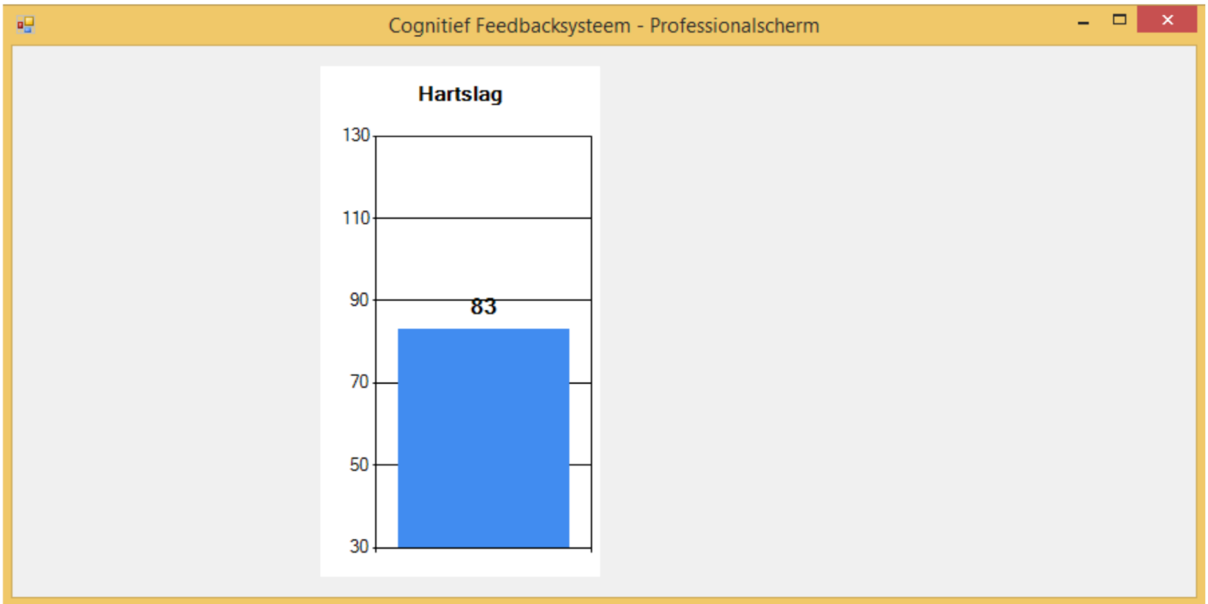


Figure 5: This Figure shows how biofeedback was presented to participants in experiment 1.

high.

Questionnaires After every scenario, the subjective data required to calibrate the COPE-model was collected. The subjective experience stress, the subjective task demand and the participant's appraisal were measured by asking people how they experienced those factors right after performing the task.

3.1.5 Procedure

Participants individually conducted the tasks in this experiment. The experiment took participants between 90 to 120 minutes, depending how fast the participants read through the instructions and questionnaires. There were sixteen scenarios that lasted for about three minutes each, during which participants had to fight all fires that appeared. All participants had to perform the same sixteen scenarios but in a different order. First, the order of scenarios 1- 8 was randomized, then the order of scenarios 9-16 was randomized. Scenarios 1-8 were used for the first half of the experiment and in the second half of the experiment, participants performed scenarios 9-16.

For every participant, the following procedure was followed:

- Experimenters determined the order in which the scenarios would be presented to participants and whether this participant would receive feedback in the first or second half of the experiment.
- Participant arrived and was asked to turn off his or her phone and to put on the heart rate monitor.
- Experimenters started CFSystem and checked if the heart rate monitor worked, while participant read instructions.
- Participant was asked whether instructions were clear and participant could ask questions if there were any.
- Participant was given the first decision tree and performed the practice trial.

- If necessary, the feedback screen was turned on.
- Experimenters started first scenario and session of the CFSystem simultaneously.
- Participant performed experimental task.
- Participant filled in questionnaire about scenario while experimenters set up the next scenario.
- Participant performed three more scenarios, filling in the questionnaire after each scenario.
- Participant was given new decision tree to prevent strong learning effect.
- Participant performed four more scenarios, filling in the questionnaire after each scenario.
- Feedback screen was turned off or on and participant was given new decision tree.
- Participant performed four more scenarios, filling in the questionnaire after each scenario.
- Participant was given new decision tree.
- Participant performed four more scenarios, filling in the questionnaire after each scenario.
- Participant filled in a survey on the experiment in general and provided demographic information.
- Participant was asked to take the heart rate monitor off and was debriefed.

All four -Dutch- decision trees that were used can be found in appendix B.

3.1.6 Data analysis

The stress-level-questionnaire data was used to find out whether performing the task was stressful for participants and to select scenarios that were suitable for experiment 2. Because scenarios had been presented to participants in a randomized order, the questionnaire data was first organized in a way that data of all participants was in the same order. Then, all five-point scale answers -do not agree at all, do not agree, neutral, agree a little, agree very much- given by participants were converted into numbers from one to five. For every scenario the median score was calculated, using data of all participants. Data of scenarios with the same parameters was then combined by averaging their medians. Then scenarios were ranked from most stressful or highest task demand to least stressful or lowest task demand. Table 3 shows these rankings.

A Generalized Linear Model was used to build predictive models of the selected scenarios. The heart rate and heart rate variability data, combined with threat and challenge-values, task demand values and the number of mistakes of different scenarios were used to calibrate the COPE-model. For the prediction of performance, a linear regression was used. For the error models, logistic linear regressions were used.

3.2 Results

Scenarios were ranked on how stressful participants had rated them. The second column of Table 3 shows the ranking from the most stressful scenario to the least stressful scenario. Columns three, four and five show rankings of scenarios from highest to lowest subjective task demand.

Table 3 shows that scenario 6 has the highest subjective task load, but was one of the less stressful scenarios. Because participants had saved fewer lives in scenario 6 than in other scenarios scenario 6 was considered to be too hard, causing participants to just give up and stress less. Thus, scenario 6 was not used for experiment 2. Scenarios 8 and 16, 2 and 10, 4 and

Table 3: Scenarios ranked by Median.

Ranking	Stress	TO	LIP	TSS
Most stressful/highest task load	8	6	6	6
	2	8	8	4
	4	4	4	2
	5	5	5	1
	6	1	2	5
	1	7	1	7
	7	2	7	3
Least stressful/lowest task load	3	3	3	8

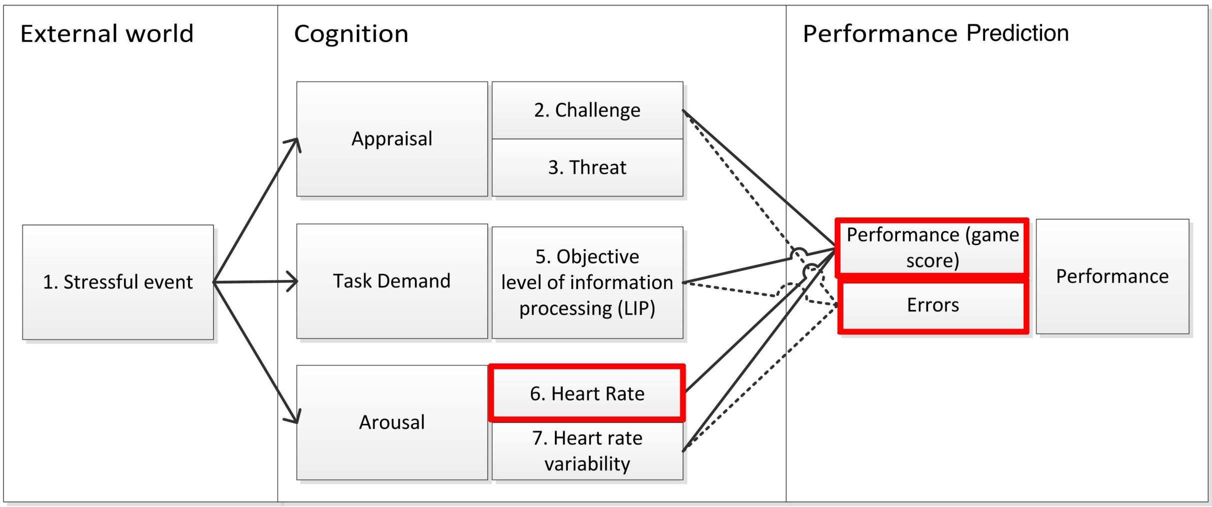


Figure 6: This Figure shows the calibrated COPE-model for experiment 2. Elements of the model that will be shown to participants as feedback, have a red outlining.

12, and 5 and 15 were the most stressful and scored high on task demand, so these scenarios were selected for experiment 2.

The COPE-model was calibrated with the data of the selected scenarios. With standardized performance scores (standardized per participant) the model could predict performance ($F(5, 24.44) = 24.227$, $p < 0{,}01$). Table 4, 5, 6 and 7 shows which measures were used to predict performance and which measures were used to predict errors and the coefficients of those measures. Figure 6 shows the calibrated COPE-model.

Communication and task allocation errors had been registered, so predictive models were made for these types of errors. No speed- or planning errors were registered, which made it impossible to create predictive models for these kinds of errors. To calibrate the COPE-model,

Table 4: Significant variables were used to predict performance.

	coefficient	std. error	t	Sign
Intercept	-0.028	0.081	-0.349	0.73
HR	**-0.003**	0.001	-2.855	**<0.01**
HRV	**0.013**	0.003	3.951	**<0.01**
Challenge	**0.033**	0.009	3.539	**<0.01**
Threat	0.009	0.006	1.570	0.12
LIP	**0.048**	0.008	6.364	**<0.05**

Table 5: Significant variables were used to predict the total errors. Errors are weighted with 25:75.

	coefficient	std. error	t	Sign
Intercept	3.233	0.461	7.016	**<0.01**
HR	-0.005	0.004	-1.271	0.20
HRV	**-0.045**	0.009	-5.176	**<0.01**
Challenge	**0.215**	0.034	6.272	**<0.01**
Threat	0.016	0.019	0.881	0.38
LIP	**-0.191**	0.029	-6.698	**<0.01**

Table 6: All variables were used to predict communication errors. These errors are weighted with 15:85.

	coefficient	std. error	t	Sign
Intercept	**9.384**	1.783	5.264	**<0.01**
HR	**-0.026**	0.009	-3.013	**<0.01**
HRV	**-0.093**	0.011	-8.818	**<0.01**
Challenge	**-0.451**	0.057	-7.904	**<0.01**
Threat	**-0.192**	0.031	-6.109	**<0.01**
LIP	**-0.389**	0.053	-7.351	**<0.01**

weightings had to be used on the data, because the number of errors made was relatively small.

The total error cases were weighted with the ratio of 25:75. By multiplying the data with a different ratio, the ratio of errors versus no errors could be distributed more evenly. For the communication and task allocation errors a ratio of 15:85 was used. Total errors could be predicted based on Heart Rate Variability, challenge, threat and Level of Information Processing ($F(5, 24.44) = 17.455$, $P < 0{,}01$). Coefficients of different variables are shown in Table 5.

Variables and coefficients to predict communication and task allocation errors are shown in Table 6 and Table 7. Communication errors were be predicted using all the variables ($F(5, 14.744) = 52.566$, $P<0{,}01$) and task allocation errors were predicted out of Heart Rate Variability, challenge, threat and Level of Information Processing ($F(5, 14.884) = 51.775$, $P<0{,}01$). More information on how the COPE-model was calibrated, can be found in the paper of Cohen, Zuiderduin, Kaiser, Brinkman and Neerincx [5].

Table 7: Significant variables were used to predict task allocation errors. These errors are weighted with 15:85.

	coefficient	std. error	t	Sign
Intercept	1.626	0.912	1.783	0.08
HR	0.000	0.004	0.078	0.94
HRV	0.725	**0.269**	2.695	**<0.05**
Challenge	**0.609**	0.043	14.044	**<0.05**
Threat	**0.124**	0.023	5.449	**<0.05**
LIP	**-0.098**	0.035	-2.770	**<0.05**

3.3 Discussion

Participants performed sixteen scenarios with eight different combinations of parameters. Scenarios 2,4,5 and 8 were selected and Table 2 shows the combinations of parameters of the selected scenarios.

This does not mean that these combinations of parameters will evoke the most stress in all situations, it merely means that the scenarios generated with these parameters evoked the most stress in participants.

Scenarios for experiment two were created by putting scenario 2 behind scenario 10, scenario 4 behind scenario 12, scenario 5 behind scenario 13 and scenario 8 behind scenario 16. The other four scenarios were created by putting scenario 10 behind scenario 2, scenario 12 behind scenario 4, scenario 13 behind scenario 5 and scenario 16 behind scenario 8. To make sure all scenarios caused the same amount of stress, small changes were made to the scenarios. For example, a fire was added to the combination of two scenarios to prevent changes in time-pressure.

4 Experiment 2

4.1 Method

4.1.1 Ethics statement

The study was approved by the ethics committee of TNO Soesterberg. Participants were aware of the voluntarily nature of participation and they were fully debriefed about the purpose of the experiment. All participants signed a consent form and were at least 18 years old. The participant whose picture is shown, gave permission to use the photo.

4.1.2 Participants

Twenty-nine participants who were between 18 and 34 years old, with an average of 25,4 years old, participated in this experiment. Fifteen of those participants were male and one of the participants smoked regularly. All participants were naive with respect to the purpose of the experiment. People who had participated in experiment 1, were excluded of participating in experiment 2. Participants were compensated with 25 euros plus travel expenses and a bonus of 20 euros was awarded to the participant with the highest score.

Due to technical difficulties, the questionnaire data of one person was lost and the heart-rate data of four participants was not measured properly. Because feedback was based on heart rate, without properly measured heart rate, participants did not receive proper feedback. Unreliable feedback was expected to influence behavior of participants differently than reliable feedback. Participants who were excluded from heart rate analysis, were also excluded from performance-analysis.

Using SPSS, outliers of heart rate data were found and discarded. Heart rates that were more than 2.5 standard deviations away from the mean were regarded as outliers and discarded. If more than 25% of the heart rate data of a participant was discarded, this participant's heart rate data was excluded from analysis. Based on these criteria, four participants were excluded from heart rate and performance analysis. One participant was excluded from the usability-analysis. For the usability-analysis, the final sample included twenty-eight participants, of which fourteen were male. For the heart rate and performance-analysis, the final sample included twenty-five participants, of which fourteen were male.

4.1.3 Experimental design

The design of this experiment was also a full-factorial within-subject design, which means that all participants were exposed to all conditions and the conditions consisted of all possible combinations of feedback.

The independent variables of this experiment were:

- biofeedback
- performance-prediction

Table 8: Set up of different conditions.

Feedback condition	Heart rate	Performance prediction	Error prediction
Control	No	No	No
HR	Yes	No	No
Pf	No	Yes	No
Er	No	No	Yes
HRPf	Yes	Yes	No
HREr	Yes	No	Yes
PfEr	No	Yes	Yes
HRPfEr	Yes	Yes	Yes

- error-prediction

The dependent variables of this experiment were:

- performance
- stress-levels or heart rate
- System Usability Scale-score (subjective variable, measured using a questionnaire)
- Most favorite (combination of) type(s) of feedback (subjective variable, measured using a questionnaire)
- Least favorite (combination of) type(s) of feedback (subjective variable, measured using a questionnaire)

4.1.4 Set up

Apparatus and materials For the first four participants, the set-up of experiment 2 was the same as the set-up of experiment 1. After losing questionnaire-data due to connection problems, the questionnaire laptop was replaced by an a4-format printed version of the questionnaire. Figure 7 shows the set-up of experiment 2.

Scenarios There were eight equally stressful scenarios that were performed in the same order by all participants. Scenarios had a length of five minutes.

Feedback There were three types of feedback that participants could be presented with: biofeedback, a performance-prediction and an error-prediction. Figure 2 shows what the three types of feedback look like. Participants had to perform the task with all seven possible combinations of these types of feedback and in the control condition there was no feedback. Table 8 shows all different feedback-conditions. What the conditions looked like to participants is shown in Appendix C. Before the experiment, participants were told how they could interpret and benefit from the different types of feedback.

Participants were presented with one of eight conditions per scenario. The order of the conditions was balanced using a Latin square to prevent the order of the conditions from influencing results. This means that there were eight different orders, and every condition was at every position (scenarios one to eight) once. Table 9 shows the eight orders of feedback that were used.

Questionnaires Between trials, participants were asked to judge the usability of the feedback they were presented with in the previous scenario by filling in the System Usability Scale (SUS). After performing all eight scenarios, participants were asked to choose a most and a least favorite type of feedback and explain why they chose those conditions.

Figure 7: This Figure shows the set-up of experiment 2. The participant had to handle fires on the screen on the left, while feedback was presented on the screen on the right. The feedback was controlled by the laptop in the front. The feedback shows that this is condition HrPfEr.

Table 9: Orders of different conditions.

Order	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6	Scenario 7	Scenario 8
1	PfEr	Control	Er	HRPf	HR	HREr	Pf	HRPfEr
2	Control	Pf	HRPf	HR	HREr	PfEr	HRPfEr	Er
3	HRPf	PfEr	HR	Control	HRPfEr	Pf	Er	HREr
4	HREr	HRPfEr	Pf	PfEr	Control	Er	HR	HRPf
5	Er	HREr	PfEr	HRPfEr	Pf	HRPf	Control	HR
6	Pf	HRPf	HRPfEr	Er	PfEr	HR	HREr	Control
7	HR	Er	Control	HREr	HRPf	HRPfEr	PfEr	Pf
8	HRPfEr	HR	HREr	Pf	Er	Control	HRPf	PfEr

4.1.5 Procedure

The procedure of this experiment was similar to the procedure of experiment 1. Participants also participated on their own and one at a time. The experiment took about 90 minutes.

For every participant, the following procedure was followed:

- Experimenters determined which feedback-order would be used.
- Participant arrived and was asked to turn off his or her phone and to put on the heart rate monitor.
- Experimenters started CFSystem and checked if the heart rate monitor worked, while the participant read instructions and signed a consent form.
- Participant was asked whether instructions were clear and participant could questions if there were any.
- Participant was asked to sit still and watch a relaxing video for three minutes in order to perform a baseline measurement.
- Participant was given the first decision tree and started the tutorial to practice performing all aspects of the task without time pressure.
- Participant performed the practice trial.
- Feedback screen was turned on.
- Experimenters started scenario 1, the data-recording session of the CFSystem and the first feedback condition simultaneously.
- Participant performed scenario 1 of experimental task.
- Participant filled in questionnaire about scenario while experimenters set up the next scenario.
- The next feedback condition was started by the experimenters simultaneously with the next task.
- Participant performed another scenario, filling in the questionnaire afterwards.
- Participant was given new decision tree.
- Participant performed two more scenarios, filling in the questionnaire after each scenario and receiving different feedback with each scenario.
- Participant was given new decision tree.

Table 10: Number of participants that participated in a certain order and number of participants of a certain order that were analysed.

Order	Total	Usability-analysis	Performance and Heart Rate analysis
1	5	5	4
2	5	4	3
3	3	3	3
4	3	3	3
5	3	3	3
6	3	3	3
7	4	4	4
8	3	3	2
Total	29	28	25

- Participant performed two more scenarios, filling in the questionnaire after each scenario and receiving different feedback with each scenario.
- Participant was given new decision tree.
- Participant performed two more scenarios, filling in the questionnaire after each scenario and receiving different feedback with each scenario.
- Participant provided demographic information and chose favorite and least favorite type or combination of types of feedback.
- Participant was asked to take the heart rate monitor off and was done with the experiment.

4.1.6 Data analysis

Using SPSS, outliers of heart rate data – heart rate measures that were removed more than 2.5 standard deviations from the mean - were found and discarded. After filtering, data was organized by feedback conditions. To determine which feedback is the most suitable for helping stressed users, performance and heart rate of participants of different feedback conditions were compared. System Usability Scale (SUS) scores participants gave different feedback conditions were also analysed.

To compare heart rates of different participants with different feedback, the median was used. The median was calculated for every condition for every participant and corrected using the median of the baseline measurement The differences between the median of the baseline and the median of conditions were compared with a repeated measures analysis. The SUS-scores and the relative performance, the number of lives saved divided by the number of lives that could have been saved in a condition, were also calculated and compared with a repeated measures analysis. SPSS was used to do the repeated measures analysis with two-leveled factors biofeedback, performance-prediction and error-prediction. Two levels were whether a type of feedback was shown or not shown.

When the repeated measure analysis showed significance, SPSS was used to do paired-sample t-tests to find out which types of feedback were significantly different from which other types of feedback.

5 Results

Because data of some participants was excluded from analysis, conditions were not perfectly balanced. Table 10 shows the number of participants that was analysed per order.

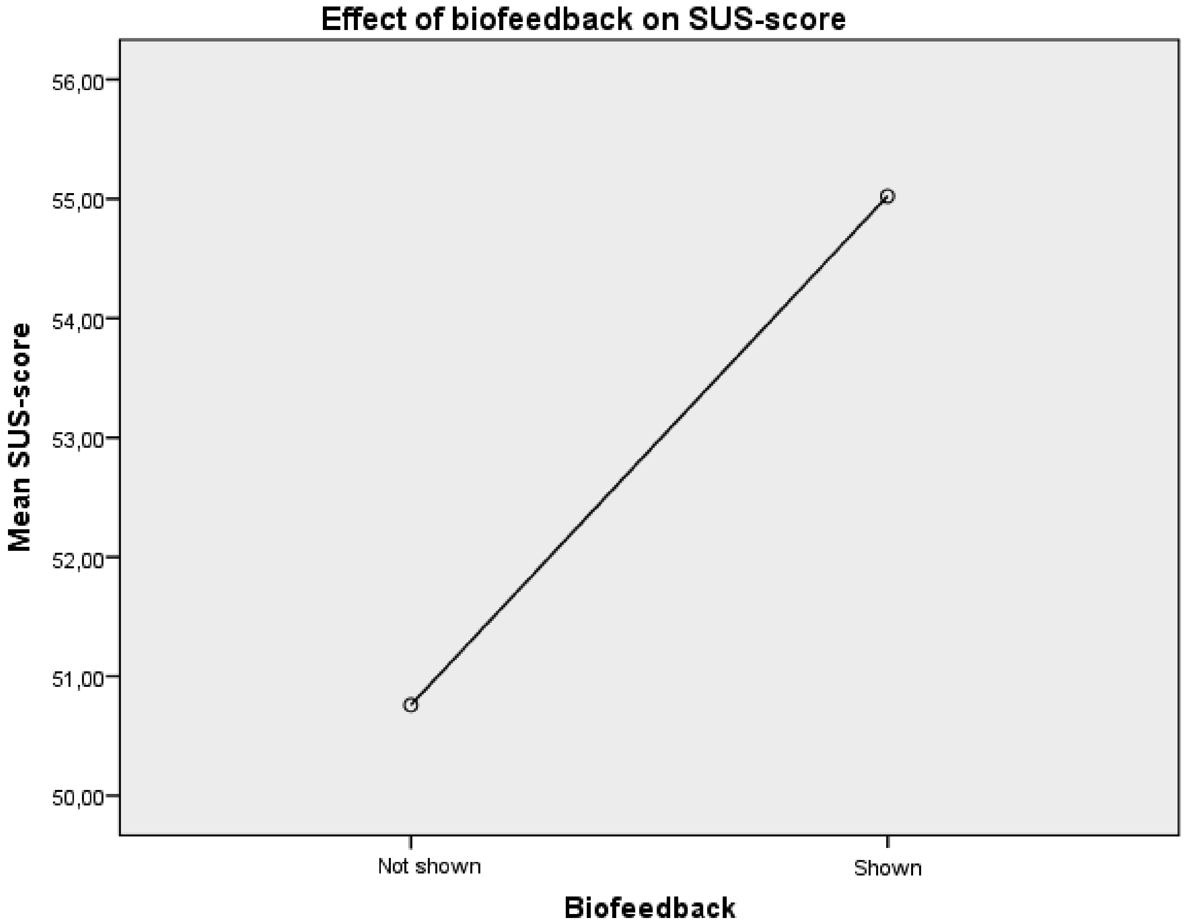


Figure 8: This graph shows the influence of biofeedback on the SUS-score.

5.1 Stress

Repeated measure analyses showed no significant main effects or interaction effects of error-prediction, performance-prediction and biofeedback (all $F(1,24)>0,007$,all $p>0,388$) on heart rate (all $F>0,007$, all $p>0,388$).

5.2 Usability

For the System Usability Scale, significant differences ($p<0,05$) were demonstrated between feedback conditions. A main effect of biofeedback on the System Usabilty Scale-score was demonstrated ($F(1,27)=7,81$, $p<0,01$) and a significant interaction of error-prediction and biofeedback ($F(1,27)=9,95$, $p<0,01$).

Figure 8 shows there is a positive influence of showing biofeedback to participants; When biofeedback was shown, participants gave the condition a higher SUS-score. However, when the error-prediction was shown in combination with biofeedback, participants gave the condition a significantly lower SUS-score. Figure 9 shows this interaction effect.

5.3 Performance

Providing participants with feedback has also shown a significant effect on performance. No main effects of types of feedback were found, but there was a significant interaction effect of error-prediction, performance-prediction and biofeedback ($F(1,24)=4,31$, $p=0,049$); Participants performance improved significantly when feedback was shown ($M=0,23$, $SE=0,06$), compared

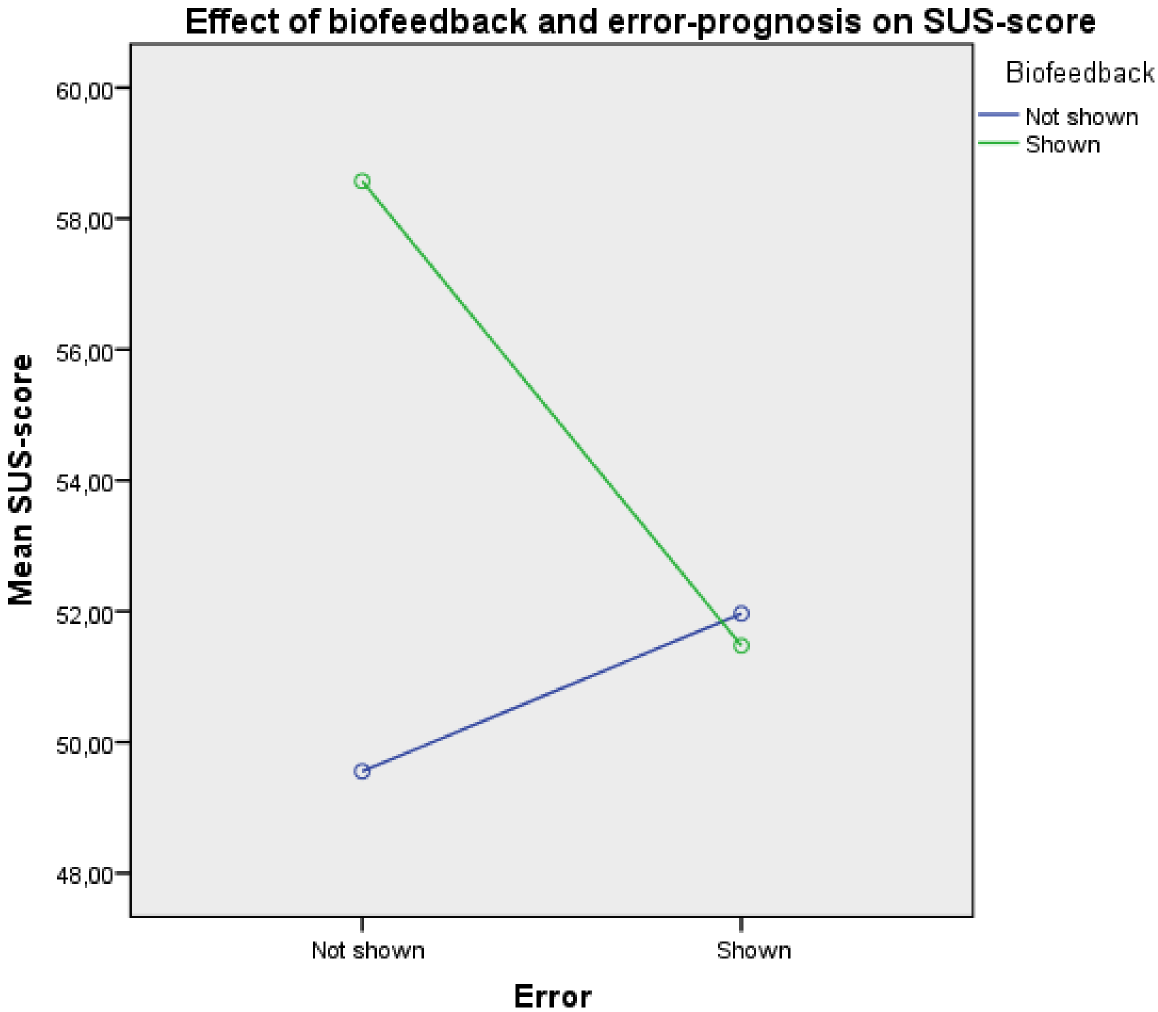
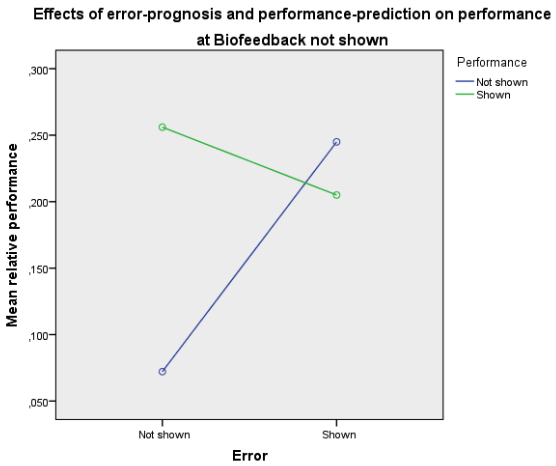
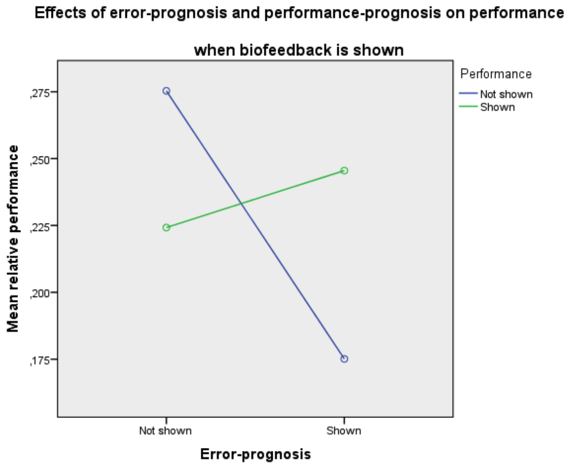


Figure 9: This graph shows the influence of biofeedback combined with the error-prediction on the SUS-score.

to performance with no feedback (M=0,07, SE=0,10), t(24)=2,18, p = 0,39. Figure 10 shows influences of different types and combinations of feedback on performance.



(a) Without biofeedback



(b) With biofeedback

Figure 10: These graphs show the influence of all different types and their combinations on performance of participants.

Besides scoring the usability of different feedback conditions, participants also had to choose a favorite and least favorite feedback condition. Figure 11 shows how many participants chose a certain type of feedback as their favorite and how many times they chose a certain type of feedback as their least favorite. Three of the 28 participants whose data was analysed, chose the same condition both as their most and least favorite type of feedback. Because these participants did not understand the questions, their data is not displayed in the graph. Figure 11 shows that participants preferred very different feedback conditions.

6 Discussion

To help emergency workers make better decisions in stressful situations, an experiment was set up to research if feedback can improve performance under stress during training. The main question was whether some types of feedback were able to improve performance more effectively then other types of feedback.

6.1 Explanation of results

No differences in heart rate were demonstrated between conditions, which is remarkable, because Meehan, Insko and Whitton [20] showed that changes in heart rate were the best physiological measure of stress. It is possible that the task did not cause enough stress, but this is implausible because participants reported experiencing stress. It can also be speculated that biofeedback reassured participants, enabling them to stay more relaxed or that feedback enabled participants to control their physiological stress reactions.

Results show that biofeedback improved usability, except when biofeedback was shown in combination with an error-prediction. The combination of the two types of feedback caused a significant decrease of SUS-scores. This implies that the usability of the error-condition should be improved and/or that presenting the error-combination together with biofeedback provided users with too much information.

Although participants were presented with more information when all three types of feedback were provided, this combination of feedback did not receive a significantly lower score than other combinations of feedback. Moreover, the combination of heart rate, performance-prediction

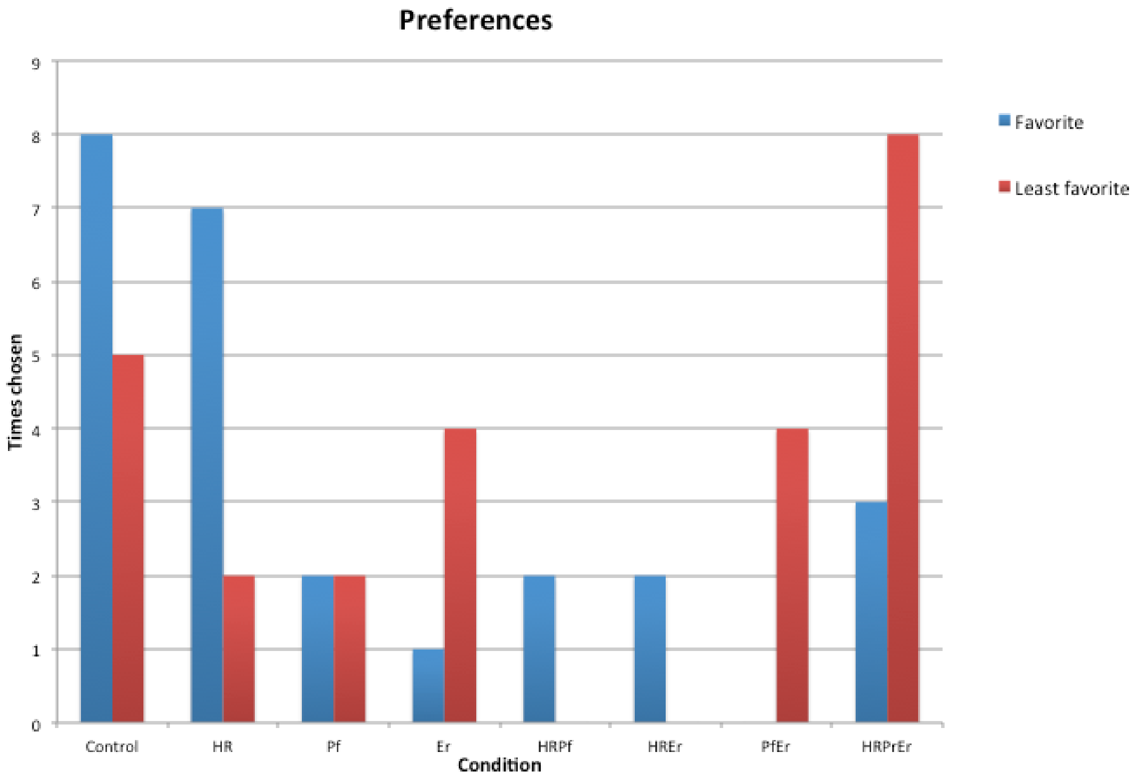


Figure 11: This graph shows how many participants chose feedback conditions as their favorite or least favorite feedback condition.

and error-prediction actually helped participants make better decisions. The hypothesis that feedback helps participants improve their performance has been confirmed by the findings that performance of participants improved when they were shown all three types of feedback and participants performed significantly better when feedback was shown in comparison to conditions where participants did not receive any feedback. The best type of feedback to help participants improve their performance is the combination of all three kinds of feedback.

Although results show that the CFSystem was helpful to participants, some improvements can be made. System Usability Scale-scores are between 0 and 100 and SUS-scores of 68 and up are considered good [3]. The biofeedback-condition scored the highest SUS-score (60,7), which means that usability of all types of feedback could be improved. Usability can change how participants perceive the quality of information [6], and usability can improve task completion times [26]. Improving usability is therefore expected to help participants trust the feedback more and interpret feedback faster. Stressed people have been demonstrated to be less able to learn from receiving feedback while performing a task, compared to relaxed people [27], so to make learning from feedback easier for participants it is important to provide feedback with good usability [6] [26].

6.2 Recommendations Cognitive Feedback System

Comments from participants imply that more instructions on how to interpret and use the feedback are also speculated to improve usability. Allowing participants to use the (different types of) feedback with a less stressful task before using the feedback with a stressful task, could make using the feedback easier for participants.

Participants also reported problems with interpreting the error-prediction, because it provided too much information and thus was hard to interpret. A prediction of the chance of making a mistake in general is speculated to be easier to interpret, because the amount of information participants have to process would be reduced.

Changing the lay-out of the feedback was another suggestion of participants to improve user-experience of the CFSystem. One of the reported problems of the error-prediction was that the bar that displayed prediction of speed-errors was red. People associate red with danger and the color red has been demonstrated to impair performance [8], so this color should be not be used for the feedback-graphs. Another option is to use colors of the graphs. If the colors of the graphs indicate 'good' or 'bad' participants can quickly see how they are predicted to perform. For example, the color of the performance-prediction bar can be green (quieting and agreeable, [8]) when predicted performance is 7 or higher, yellow or orange between a 5 and 7 and red when the performance-prediction of a participant is 5 or lower. One might speculate that giving the sizes of the bars the same meaning, would also make the feedback easier to interpret. Instead of the performance-prediction being better when the bar is high while error-predictions are worse when they are high, all bar-sizes should mean 'better' when they are higher.

A proposal to help participants keep dividing their attention between their task and looking at the feedback-screen to a minimum, which is expected to lower cognitive task load by lowering Task Set Switches [17] audible feedback could be added to the CFSystem. For example, when the heart rate or error-prediction of a participant exceeds a certain value, this could trigger a sound. When they hear the sound, participants can check the feedback screen to see which value is too high or too low, or different sounds for different types of feedback can be used. This is not possible when there is too much noise from the task itself.

It is speculated that usability of the biofeedback-condition would improve if the CFSystem showed the difference between the current heart rate and the baseline heart rate of the participant. This would enable participants to determine more easily when their heart rate is high and when their heart rate is low.

Because participants preferred and disliked very different types of feedback, providing them with the option to only show the types of feedback they prefer is also speculated to improve usability. This would prevent the CFSystem from providing information that the user does not want and allow the user to focus their attention on interpreting feedback he or she wants to use.

6.3 Intelligent Decision Aids

In general, IDA's should be easy to interpret and using them should not take away (too much) attention from the task itself, to help users improve their performance [6] [13] [26]. Furthermore, providing all three types of feedback to participants while they had to perform a stressful task, might have caused cognitive overload and reduced the ability of participants to learn from the feedback [29]. However, results of this experiment show that feedback with highest usability-score did not improve performance, while feedback that provided the most information did improve performance.

A possible explanation is a difference in the way participants processed different combinations of feedback. One might speculate that when one or two types of feedback were provided, participants explicitly – controlled and rule-based cognitive processes – process the information that was provided by the CFSystem. This caused no significant improvements in performance. However, participants are not expected to be able to process all information provided by the CFSystem and perform the stressful task, when they were provided with all three types of feedback. Thus, participants were expected to benefit less from the combination of all three types of feedback. Although unclear what the exact roles of the two processes are, deliberate, goal-based decision making consists of both explicit and an implicit processes – automatic and involving emotional reactions [12]. Some theoretical models have been developed, but little applied research has been done on the role of implicit and explicit processes and decision making, and there is not much scientific work on implicit decision making, or intuition [16]. This is probably caused by the idea that intuition is the opposite of rational and as mentioned earlier, rationality is often seen as the only way to make a good decision [9]. The finding that performance of participants improved when the CFSystem provided them with all possible types of feedback, might be an indication that users processed the information provided by the CFSystem implicitly. This implies that implicit processes may play a very important part in decision making or at least in improving decision making under stress, and that participants benefit more from implicitly processed feedback. Because implicit processes take up less working memory capacities than explicit processes or even or no working memory capacities [12], implicitly processed feedback can support users without taking away much capacities necessary for performing the task. However, more research should be done to gather more knowledge about implicit reasoning and the role of implicit processes in decision making.

It is also possible that feedback in general reduces uncertainty. The combination of biofeedback, performance-prediction and the error-prediction provided the most information, thus reduced uncertainty more, which could have made participants feel more certain about their performance, reduced stress and helped them improve their performance [27]. In this case, just knowing the feedback is there when participants need it, might be enough to improve performance.

Both possible explanations assume that users do not process the feedback thoroughly. This experiment therefore indicates that feedback that supports users by providing them with information on their performance does not need to be specific.

6.4 Limitations

A few things should be noted about this research. First, the findings in this experiment were based on an experiment that used a computer simulation of an emergency situation. Although the task induced stress, it is possible that rescue workers would react differently when using feedback of the CFSystem in a situation where real people are in danger. To get results that also apply to naturalistic decision making, most of the eight features of decision making in naturalistic environments Orasanu and Connolly [23] listed – ill-structured problems, uncertain dynamic environment, shifting or competing goals, action or feedback loops time stress, high stakes, multiple players and organizational goals and norms – were taken into account when designing the task.

Second, participants had a strong learning effect for performing this task. By providing participants with a new decision tree every few trials and by shuffling the order of the questions on the screen, the learning effect has been limited. By balancing the order in which participants were presented with different types of feedback, effects of the learning effect on the results were cancelled out.

Third, these results are based on short periods of time. It is possible that the feedback has stronger or different effects when participants use it for a longer period, because implicit processes are expected to be modified by slow incremental learning processes [12].

More research should be done to learn more about how stressed people process and react to feedback when they are performing a task. For example, it is unclear whether participants looked at and understood all types of feedback that were presented. Maybe they just looked at some of the types of feedback. Furthermore, experiments should be set up to research whether feedback has the same effects for users that are good at performing the task and will probably receive positive, reassuring feedback, and users that are bad at performing the task who will probably receive very different, more alarming feedback. It is possible that results of feedback for those people causes more stress which can cause participants to perform less good than they would have done without feedback. Finally, more research on implicit processes in decision making should be done.

6.5 Conclusion

Although not all hypotheses have been confirmed, and although usability of the feedback provided by the CFSystem was not very good, providing feedback did help participants improve their performance. The combination of biofeedback, performance-prediction and error-prediction was the only combination of feedback that helped participants improve their performance and was the best combination of feedback to help stressed users make better decisions. Because participants are expected to benefit more from feedback with better usability [26], usability of the feedback should be improved.

Results also implied that feedback was processed implicitly and that implicitly processed feedback was more helpful to participants than explicitly processed feedback. More research should be done on how implicit processes in decision making work and how implicit processing can be used to help people in emergency professions improve performance even more

7 Acknowledgments

Thanks to Maarten van der Smagt, my university supervisor and Mark Neerincx for the useful feedback. Many thanks to Iris Cohen, my supervisor at TNO, for all the help. Also a big thank you to Mira Kaiser, an intern at TNO, for programming the firefighting task within a very short period of time and thanks to Willem-Paul Brinkman of the TU Delft for help with the data analyses. Thanks to Bram de Beer for help with organizing the data.

References

[1] B. W. Balleine. The neural basis of choice and decision making. *The Journal of neuroscience*, 27(31):8159–8160, 2007.

[2] S. Bouchard, F. Bernier, É. Boivin, B. Morin, and G. Robillard. Using biofeedback while immersed in a stressful videogame increases the effectiveness of stress management skills in soldiers. *PloS one*, 7(4):e36169, 2012.

[3] J. Brooke. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189:194, 1996.

[4] I. Cohen, W.-P. Brinkman, and M. A. Neerincx. Predicting performance and errors in a stressful training scenario on a naval ship simulator by environmental and cognitive factors. In preparation. For more information, contact the author.

[5] I Cohen, S. M. Zuiderduin, M. Kaiser, W.-P. Brinkman, and M. A. Neerincx. What kind of feedback do people prefer when working on difficult tasks under stress? heart rate, performance prediction, error predictions or a combination of those.

[6] A. De Angeli, A. Sutcliffe, and J. Hartmann. Interaction, usability and aesthetics: what influences users' preferences? In *Proceedings of the 6th conference on Designing Interactive systems*, pages 271–280. ACM, 2006.

[7] J. Dowell. Coordination in emergency operations and the tabletop training exercise. *Le Travail Humain: A Bilingual and Multi-Disciplinary Journal in Human Factors*, 1995.

[8] A. J. Elliot, M. A. Maier, A. C. Moller, R. Friedman, and J. Meinhardt. Color and psychological functioning: the effect of red on performance attainment. *Journal of Experimental Psychology: General*, 136(1):154, 2007.

[9] C. L. Gohm, M. R. Baumann, and J. A. Sniezek. Personality in extreme situations: Thinking (or not) under acute stress. *Journal of Research in Personality*, 35(3):388–399, 2001.

[10] M. Grootjen, E. P. B. Bierman, and M. A. Neerincx. Optimizing cognitive task load in naval ship control centres: Design of an adaptive interface. In *IEA 16th World Congress on Ergonomics*, 2006.

[11] M. Grootjen, M. A. Neerincx, K. D. Stolk, J. C. M. van Weert, and E. P. B. Bierman. Design and user evaluation of an interface prototype that adapts to the operator's cognitive task load. *Proceedings of the 4th International Conference of the Augmented Cognition*, pages 97–106, 2007.

[12] R. Hastie. Problems for judgment and decision making. *Annual review of psychology*, 52(1):653–683, 2001.

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

[14] K. S. Kassam, K. Koslov, and W. B. Mendes. Decisions under distress stress profiles influence anchoring and adjustment. *Psychological Science*, 20(11):1394–1399, 2009.

[15] G. Keinan. Decision making under stress: scanning of alternatives under controllable and uncontrollable threats. *Journal of personality and social psychology*, 52(3):639, 1987.

[16] N. Khatri and H. A. Ng. The role of intuition in strategic decision making. *Human Relations*, 53(1):57–86, 2000.

[17] T. Kontogiannis and Z. Kossiavelou. Stress and team performance: principles and challenges for intelligent decision aids. *Safety Science*, 33(3):103–128, 1999.

[18] R. S. Lazarus and S. Folkman. Stress. *Appraisal, and coping*, page 456, 1984.

[19] A. J. Maule, G. R. J. Hockey, and L. Bdzola. Effects of time-pressure on decision-making under uncertainty: changes in affective state and information processing strategy. *Acta psychologica*, 104(3):283–301, 2000.

[20] M. Meehan, B. Insko, M. Whitton, and F. P. Brooks Jr. Physiological measures of presence in stressful virtual environments. In *ACM Transactions on Graphics (TOG)*, volume 21, pages 645–652. ACM, 2002.

[21] G. Meij. Sticking to plans: capacity limitation or decision-making bias? 2004.

[22] M. A. Neerincx, S. Kennedie, M. Grootjen, and F. Grootjen. Modeling the cognitive task load and performance of naval operators. In *Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience*, pages 260–269. Springer, 2009.

[23] J. Orasanu and T. Connolly. The reinvention of decision making. 1993.

[24] E. J. A. Schreuder and T. Mioch. The effect of time pressure and task completion on the occurrence of cognitive lockup. In *CEUR Proceedings 4th Workshop HCP Human Centered Processes*, 2011.

[25] M.-G. Seo and L. F. Barrett. Being emotional during decision making?good or bad? an empirical investigation. *Academy of Management Journal*, 50(4):923–940, 2007.

[26] A. Sonderegger and J. Sauer. The influence of design aesthetics in usability testing: Effects on user performance and perceived usability. *Applied ergonomics*, 41(3):403–410, 2010.

[27] K. Starcke and M. Brand. Decision making under stress: a selective review. *Neuroscience & Biobehavioral Reviews*, 36(4):1228–1248, 2012.

[28] K. Starcke, O. T. Wolf, H. J. Markowitsch, and M. Brand. Anticipatory stress influences decision making under explicit risk conditions. *Behavioral neuroscience*, 122(6):1352, 2008.

[29] J. Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2):257–285, 1988.

8 Appendix A - Dutch instructions

Experiment 1

Welkom!

Bedankt voor je deelname aan ons onderzoek. Dit blad bevat informatie over het onderzoek en instructies voor de taak. Lees deze rustig door en als je daarna nog vragen hebt, kun je die stellen voor je begint. Het lijkt veel informatie, maar schrik niet, je hoeft niet alles uit je hoofd te leren! Naast uitgebreide instructies zijn er een tutorial en een training.

Het de bedoeling dat je straks de rol van calamiteitencoördinator van een schip op je neemt. Je taak als calamiteitencoördinator bestaat uit het beschermen van een virtueel schip en zijn bemanning, door de branden die ontstaan op het schip te blussen. Jouw taak is om informatie te verzamelen over de brandjes en te bepalen wat de juiste manier is om branden te blussen.

Tijdens het uitvoeren van deze taak zal je stress worden gemeten door middel van een hartslagmeter en huidgeleidings- en spiersensors. Verder krijg je een koptelefoon op om waarschuwingssignalen te kunnen horen.

Na afloop krijg je een vragenlijst voorgelegd, om je ervaringen met de taak en het onderzoek te delen. Ook tussendoor word je af en toe gevraagd een korte vragenlijst in te vullen.

Lees de instructies op de volgende bladzijden goed door voor je begint en nogmaals bedankt voor je deelname!

Uitleg

De taak

Op het scherm zie je straks het schip afgebeeld zoals in Figuur 12. Brandjes die ontstaan worden aangeduid met een vuursymbool, met daarop meestal een timer die aangeeft hoe lang je nog hebt om de mensen te redden. Ook zijn er urgente brandjes die worden gekenmerkt door onvoorspelbaarheid. Hiervan is niet bekend hoe lang je nog hebt voor er slachtoffers vallen, maar het is korter dan bij normale branden. Ook zijn er meer mensen in gevaar, dus het is verstandig om urgente branden eerst af te handelen.

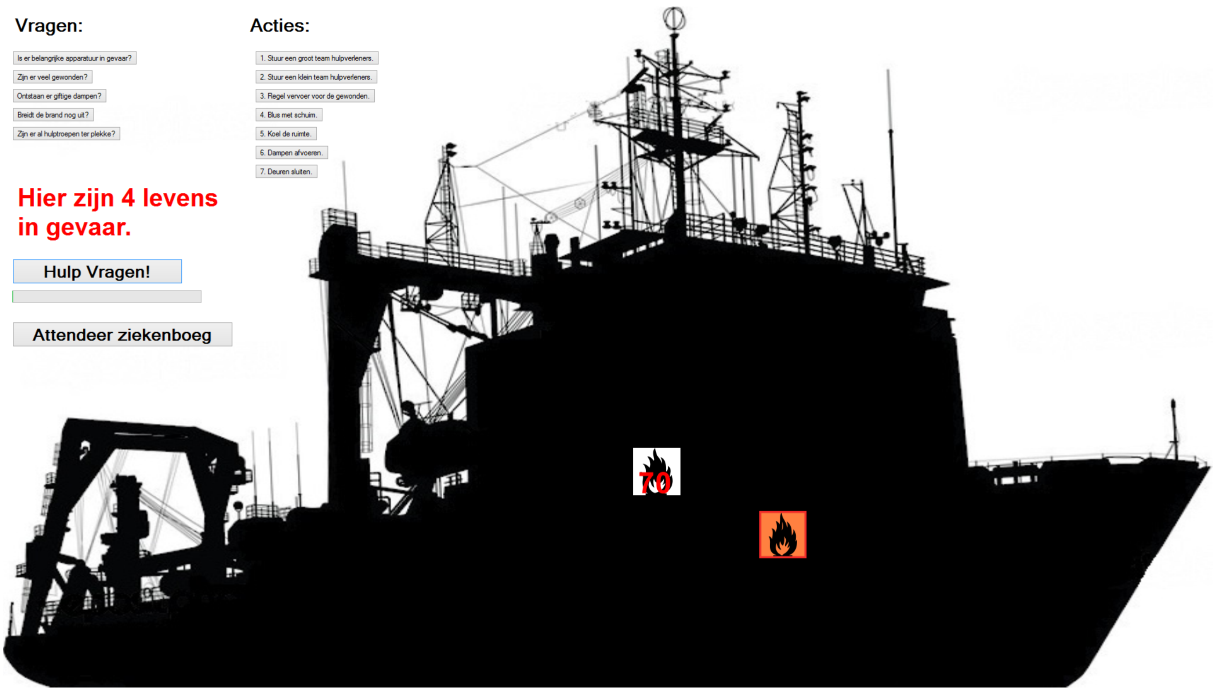


Figure 12: Zo ziet het schip er straks uit op je scherm. Het vuursymbool geeft de locatie van een brand aan. Klik er op om informatie over de brand te kunnen zien en verzamelen.

Om een brand af te handelen **begin je door er op te klikken**. Daarna verzamel je zo snel mogelijk alle benodigde informatie over een brand door de **juiste vragen te stellen** met behulp van de beslisboom. Je kunt ook alle vragen stellen zonder in de boom te kijken, maar niet alle vragen zijn altijd nodig om de juiste actie te vinden. Omdat het even duurt voor informatie beschikbaar is, kan het stellen van overbodige vragen onnodig veel tijd kosten. Wanneer alle benodigde informatie bekend is, kun je de beslisboom doorlopen en de juiste aanpak voor de brand bepalen. Het is ook mogelijk een actie te selecteren voordat alle vragen beantwoord zijn, maar de kans is klein dat je zo de goede actie kiest. Selecteer die actie door er op te klikken. Als het de juiste actie was, verdwijnt de brand van je scherm en heb je mensen gered. Als het niet de juiste actie was, verlies je alle levens die op het spel stonden en duurt het even voor je verder kunt gaan. Het verschilt per brand hoeveel mensen er in gevaar zijn en het is jouw **doel om zo veel mogelijk mensen te redden en zo min mogelijk mensen te verliezen**. Bij het bepalen hoe goed je gepresteerd hebt, worden verloren levens in mindering gebracht op geredde levens. Als je bijvoorbeeld vijf mensen hebt gered, maar ook vijf mensen bent verloren, dan heb je effectief nul mensen gered.

Omdat de taak veel van je vraagt, kan het zijn dat je af en toe het gevoel hebt dat je het niet meer aan kunt. Voor die situaties is er de mogelijkheid om het afhandelen van de taak te delegeren aan een virtuele collega. Dit doe je door op de **knop "Hulp vragen"** te klikken. Let hierbij wel op dat de collega minder ervaren is en het daarom niet in staat is om met alle

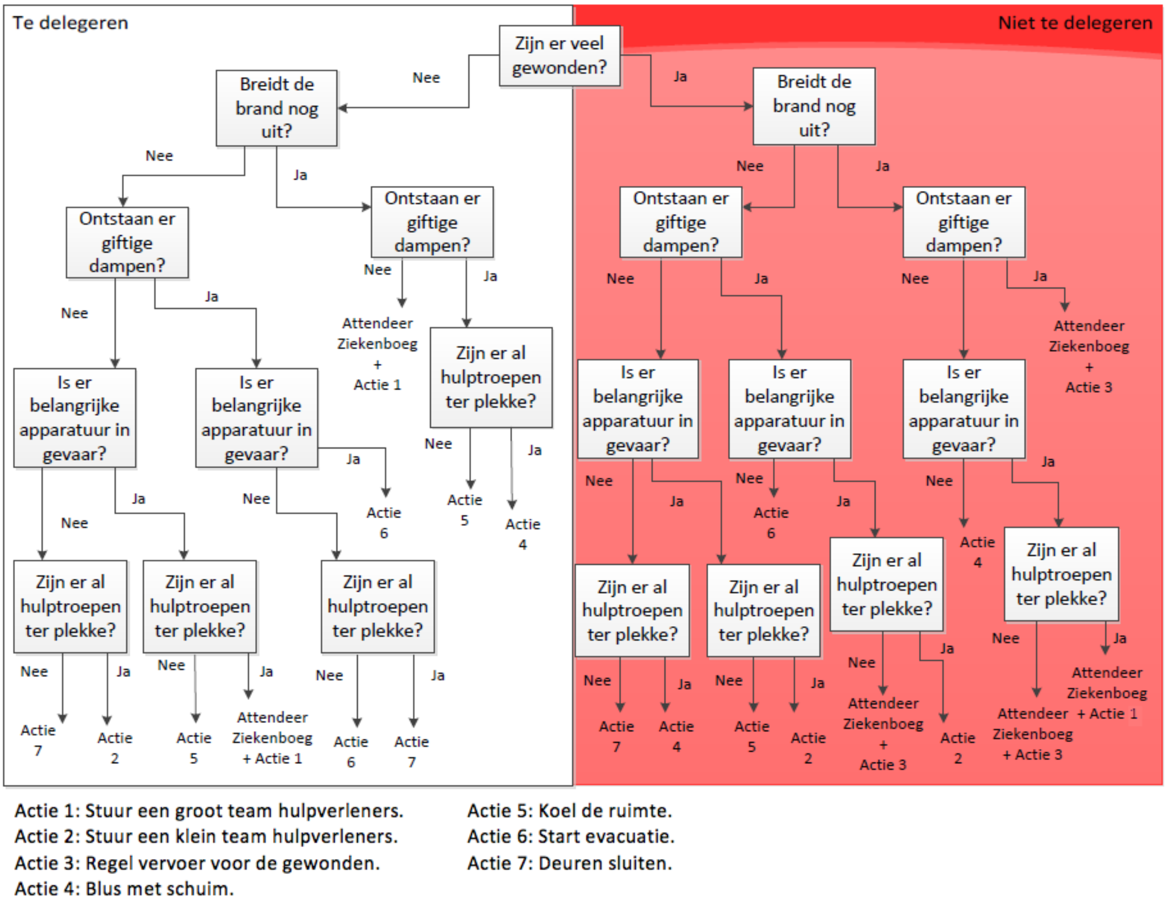


Figure 13: Schema om de actie te bepalen waarmee je een brand kunt afhandelen.

soorten branden om te gaan. **Urgente branden zijn bijvoorbeeld nooit te delegeren**. **In het gekleurde vlak in de beslisboom staan taken die ook te moeilijk zijn voor je collega**. Je kunt deze branden wel uitbesteden, maar de collega kan ze niet aan en alle levens die op het spel staan zullen verloren worden. Ook heeft de collega voldoende tijd nodig voor je opnieuw een brand door kunt geven. De collega is minder ervaren en hierdoor langzamer, dus ook bij het correct doorgeven van een brand valt er altijd een slachtoffer.

Bij bepaalde branden is het belangrijk dat jij het vervoer van de gewonden regelt. Als er veel gewonden vervoerd moeten worden, moet dat doorgegeven worden aan de ziekenboeg door op de knop "Attendeer ziekenboeg". Zij kunnen zich dan voorbereiden op de komst van de gewonden en ze beter opvangen. **In de boom staat aangegeven wanneer je de ziekenboeg moet attenderen**. Klik op de knop **voor** je een actie selecteert, anders valt er een slachtoffer.

Tenslotte wordt er bij een aantal taken een grafische weergave van je hartslag weergegeven op het tweede scherm zoals in Figuur 14. Als je hartslag erg hoog is, is dit een indicatie van stress. Als je gestrest bent, ben je meer geneigd om fouten te maken, dus probeer dit te voorkomen. Als je ziet dat je gestrest bent, probeer dan beter te letten op wat je doet om fouten te voorkomen. Gebruik dit in je voordeel.

Na het uitvoeren van een taak is er een korte vragenlijst. Na afloop van het onderzoek, dus na uitvoeren van alle taken is er nog een vragenlijst over het gehele onderzoek.

Nog even in het kort; jouw taak is om branden af te handelen. Dat doe je door de volgende stappen uit te voeren:

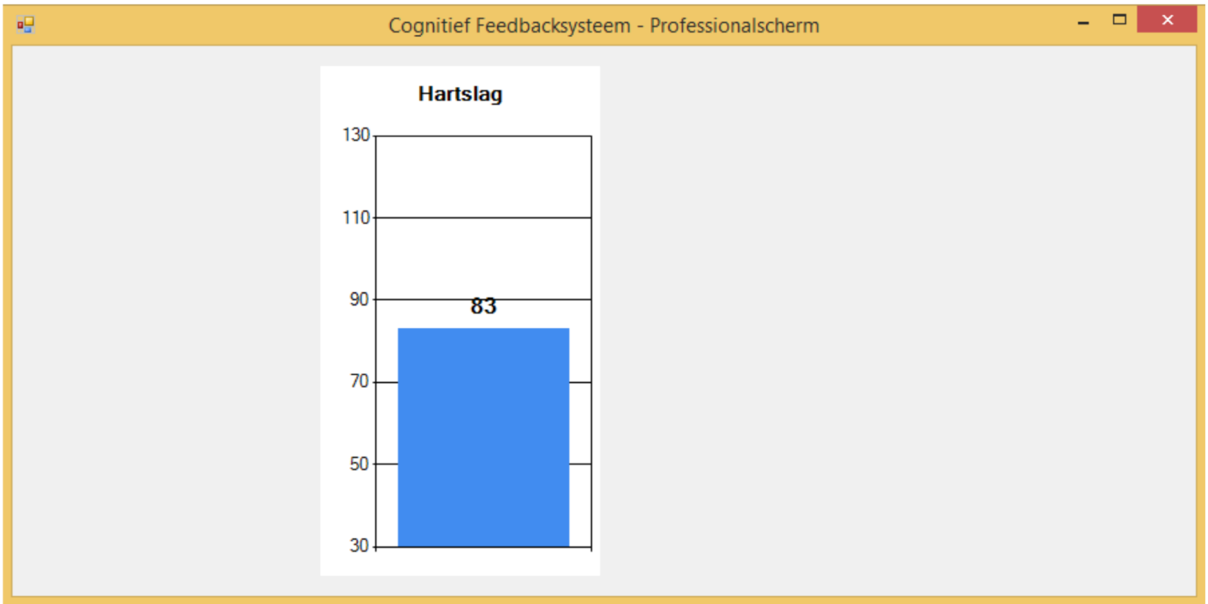


Figure 14: Ondersteuning bij de taken.

- Selecteer de brand die je wil blussen.
- Stel de vragen die nodig zijn aan de hand van de beslisboom.
- Attendeer de ziekenboeg als dat nodig is.
- Selecteer de juiste actie.

Blijf goed op de beslisboom letten om uit te vinden wat de juiste actie is en of er nog aanvullende informatie is over de brand. En onthoud dat je in bepaalde situaties hulp kunt vragen met een brand als je tijd tekort komt.

Voor je begint

Voor je begint met het uitvoeren van de taak waarbij je resultaten meetellen, zijn er een tutorial en een oefenronde, zodat je goed voorbereid begint. Natuurlijk mag je ook nog vragen stellen over de taak.

Nu mag je eerst de hartslagmeter omdoen, deze moet een beetje nat zijn en moet tegen de huid gedragen worden om goed te werken. Daarna worden door de proefleiders huidgeleidings- en bewegingssensoren opgeplakt. Het is belangrijk dat je tijdens het uitvoeren van de taak zo stil mogelijk zit om de meetapparatuur niet te verstoren of los te trekken.

Is alles duidelijk?

Experiment 2

Welkom!

Bedankt voor je deelname aan ons onderzoek. Dit blad bevat informatie over het onderzoek en instructies voor de taak. Lees deze rustig door en als je daarna nog vragen hebt, kun je die stellen voor je begint. Schrik niet van de hoeveelheid informatie, na het lezen van de instructies is er een tutorial en daarna nog een trainingsronde.

Zoals reeds is uitgelegd in het informatieformulier, is het de bedoeling dat je de rol van calamiteitencoördinator van een schip op je neemt. Je hebt de taak een virtueel schip en zijn bemanning te beschermen, door de branden die ontstaan te blussen. Tijdens deze taak krijg je verschillende soorten ondersteuning. Daarnaast zullen je stresslevels worden gemeten met behulp van een hartslagmeter en huidgeleidings- en spiersensors.

Na afloop krijg je een vragenlijst voorgelegd, om je ervaringen met de taak en de ondersteuning te delen. Ook tussendoor word je een paar keer gevraagd een korte vragenlijst in te vullen.

Lees de instructies goed door en nogmaals bedankt voor je deelname!

Uitleg

De taak

Brand

Op het scherm zie je straks het schip afgebeeld zoals in Figuur 24. Brandjes die ontstaan worden aangeduid met een vuursymbool, met daarop meestal een timer die aangeeft hoe lang je nog hebt om de mensen te redden. Ook zijn er urgente brandjes die worden aangegeven met een oranje en blauw knipperend vuursymbool, en worden gekenmerkt door onvoorspelbaarheid. Hiervan is niet bekend hoe lang je nog hebt voor er slachtoffers vallen, maar het is korter dan bij normale branden. Ook zijn er meer mensen in gevaar, dus het is verstandig om urgente branden eerst af te handelen. In Figuure 24 zien we links een normale brand waar nog 70 seconden over zijn om te blussen, en rechts een urgente brand.

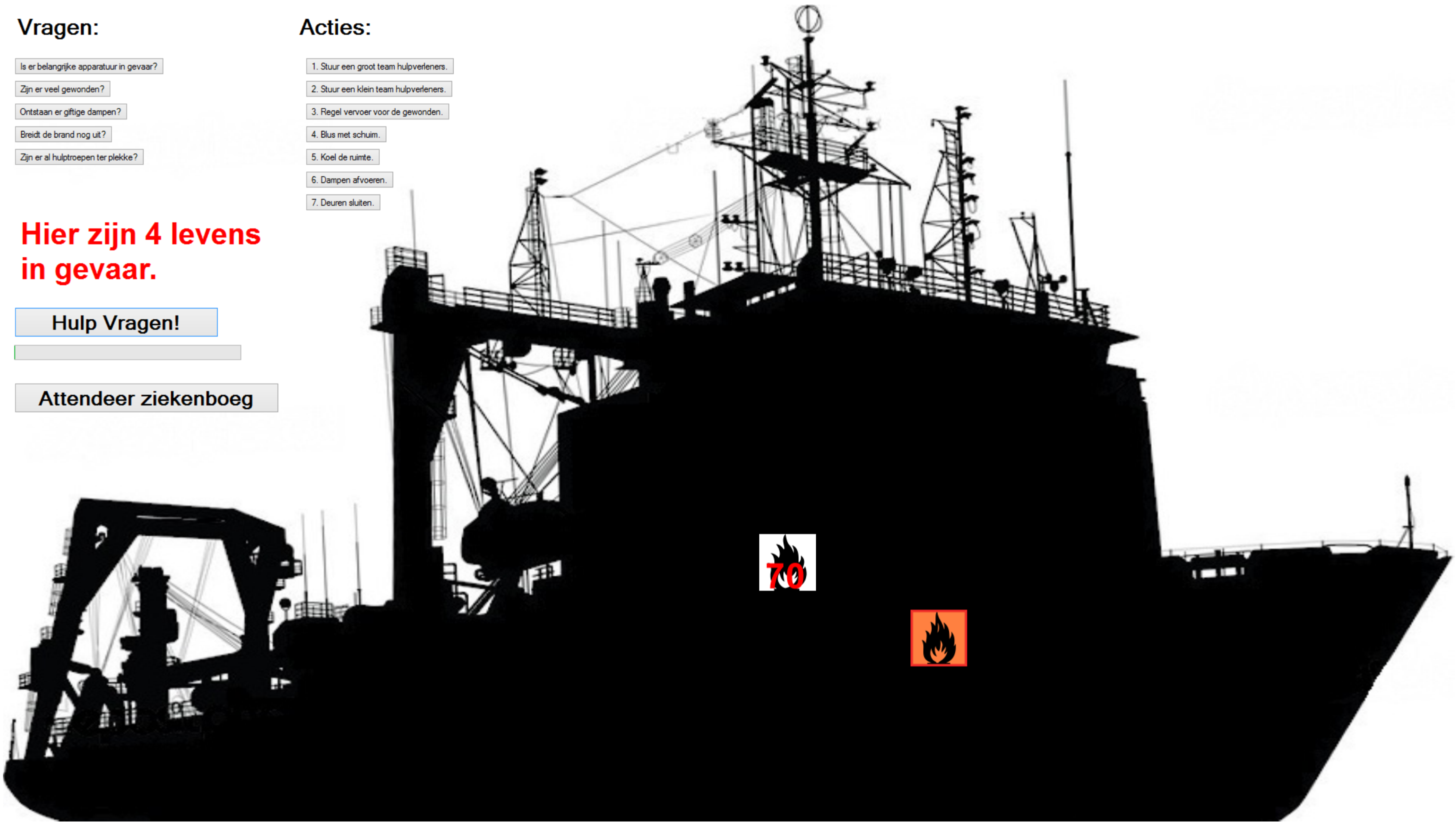


Figure 15: Zo ziet het schip er straks uit op je scherm. Klik op een vuursymbool om informatie over de brand te kunnen zien en verzamelen.

Om een brand af te handelen **begin je door er op te klikken**. Er verschijnen dan rechts boven vragen die je aan het systeem kan stellen, en acties die je kan selecteren, zoals zichtbaar in Figuur 25, en rechts boven in Figuur 24.

Beslisboom

Je kunt de brand blussen, door de juiste actie te kiezen. Je ontdekt wat de juiste actie is, aan de hand van de beslisboom, die zichtbaar is in Figuur 26, en die je er straks bij krijgt. Je kunt ook alle vragen stellen zonder in de boom te kijken, maar niet alle vragen zijn altijd nodig om de juiste actie te vinden. Omdat het even duurt voor informatie beschikbaar is en je het antwoord krijgt, kan het stellen van overbodige vragen onnodig veel tijd kosten. Wanneer alle benodigde informatie bekend is, kun je de beslisboom doorlopen en de juiste aanpak voor de brand bepalen.

Figure 16: Rechts boven verschijnen vragen (links) en acties (rechts) waar je op kunt klikken.

Selecteer die actie door er op te klikken. Als het de juiste actie was, verdwijnt de brand van je scherm en heb je alle mensen gered. Als het niet de juiste actie was, bevriest het programma even, en kun je tijdelijk niks doen. De tijd loopt wel door, en wanneer de tijd op is verlies je alle levens die in gevaar waren.

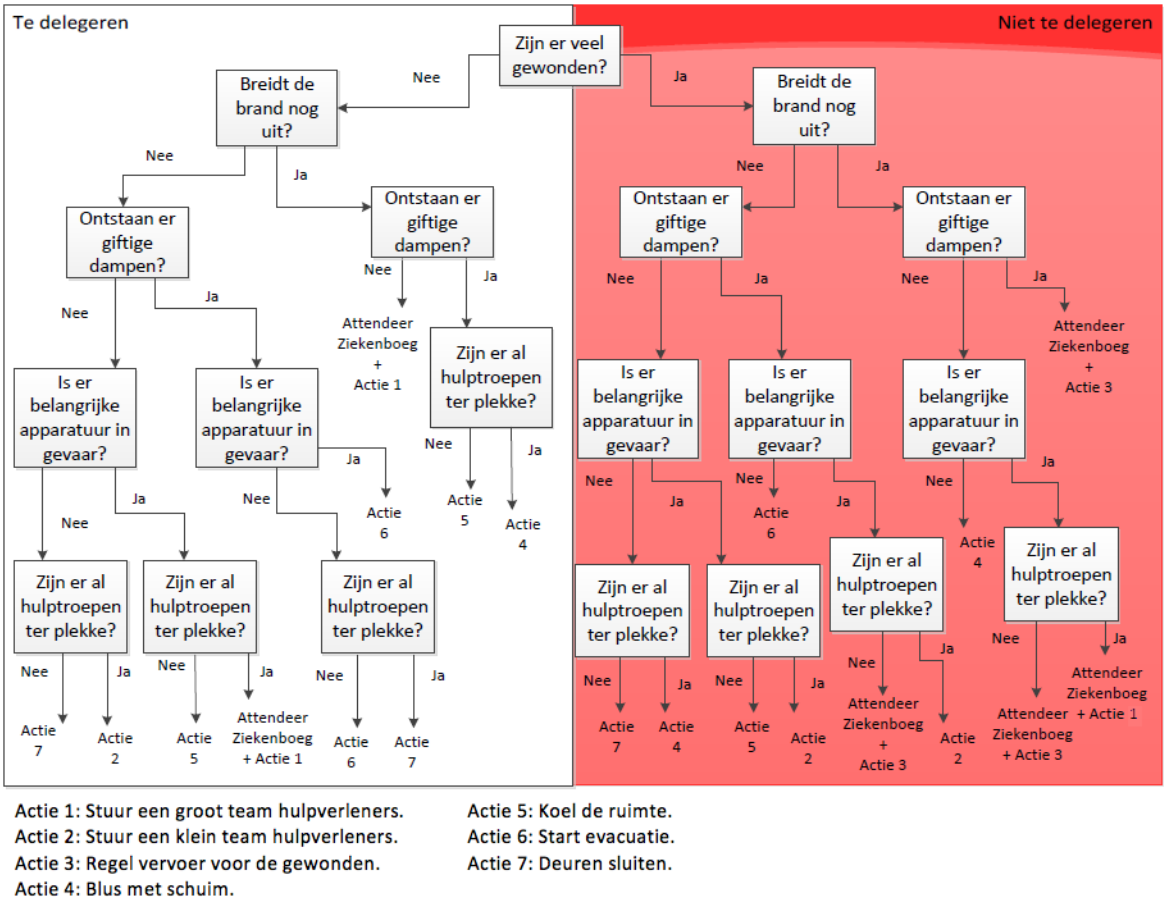


Figure 17: Schema om de actie te bepalen waarmee je een brand kunt afhandelen.

Het verschilt per brand hoeveel mensen er in gevaar zijn en het is jouw **doel om zo veel mogelijk mensen te redden en zo min mogelijk mensen te verliezen**. De deelnemer die de meeste levens redt, ontvangt een beloning van €20,-. Bij het bepalen van de deelnemer die de taak van calamiteitencoördinator het beste heeft uitgevoerd worden verloren levens in mindering gebracht op geredde levens. Als je bijvoorbeeld vijf mensen hebt gered, maar ook vijf mensen hebt verloren, dan heb je effectief nul mensen gered.

Hulp Vragen

Omdat de taak veel van je vraagt, kun je het gevoel krijgen dat je het niet meer aan kunt. Daarom is er de mogelijkheid om branden te **delegeren** aan een virtuele collega. Dit doe je door op de **knop "Hulp vragen"** te klikken, zie Figuur 27. Let hierbij wel op dat de collega minder ervaren is en daarom niet in staat is om met alle soorten branden om te gaan. **Urgente branden** zijn **nooit te delegeren**. **In het gekleurde vlak in de beslisboom staan in het rode vlak welke branden te moeilijk zijn voor je collega**. Doordat de persoon die je kan vragen om het afhandelen van een brand van je over te nemen minder ervaring heeft, is deze persoon langzamer. Hierdoor valt er **altijd minstens een dode als je een brand uitbesteedt**. Als je een te moeilijke brand uitbesteed, dus een urgente brand of een brand in het rode vlak in de boom, is de kans op meer dan een slachtoffer hoog.

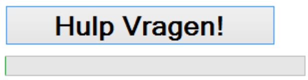


Figure 18: Deze knop gebruik je om te delegeren.

Feedback

Tenslotte krijg je op het tweede beeldscherm bij een aantal scenarios terugkoppeling van de computer. De feedback bestaat uit een grafische weergave van je hartslag, een voorspelling van je prestatie en/of een voorspelling van de fouten die je kunt maken. Je krijgt verschillende combinaties van soorten feedback te zien en je voert ook een keer de taak uit zonder feedback. Probeer de feedback in je voordeel te gebruiken. Als je hartslag verhoogd is, kan dit een indicatie zijn dat je gestresst bent. Probeer voor jezelf in te schatten of een hoge hartslag op dat moment om stress gaat of niet. Als je gestrest bent, maak je meer fouten, dus probeer dan beter te letten op wat je doet om fouten te voorkomen. Afhankelijk van de feedback die je op dat moment krijgt, kun je bij stress letten op de voorspelling van prestatie (linker kolom) of de voorspellling van fouten (rechter kolommen). Het is goed als de voorspelling van je prestatie hoog is. De voorspelling van fouten wil je juist graag zo laag mogelijk zien. Fouten waar je op geattendeerd wordt, zijn snelheidsfouten, dus te langzaam handelen, communicatiefouten, deze maak je als je vergeet de ziekenboeg te attenderen of meer vragen stelt dan nodig zijn om de brand te blussen, planningsfouten, als je een minder urgente brand of brand waar je nog meer tijd voor hebt eerst blust en taakverdelingsfouten, deze maak je als je een brand delegeert terwijl dit niet mag. In Figuur 28 zie je een voorbeeld van hoe de verschillende typen ondersteuning er uitzien.

Na afloop van een scenario word je gevraagd in een korte vragenlijst je ervaringen met de feedback te delen. Het eind van een scenario wordt duidelijk aangegeven. Na het invullen van de vragen over het afgelopen scenario staat het nieuwe scenario voor je klaar en mag je op "start" klikken om door te gaan met het volgende scenario.

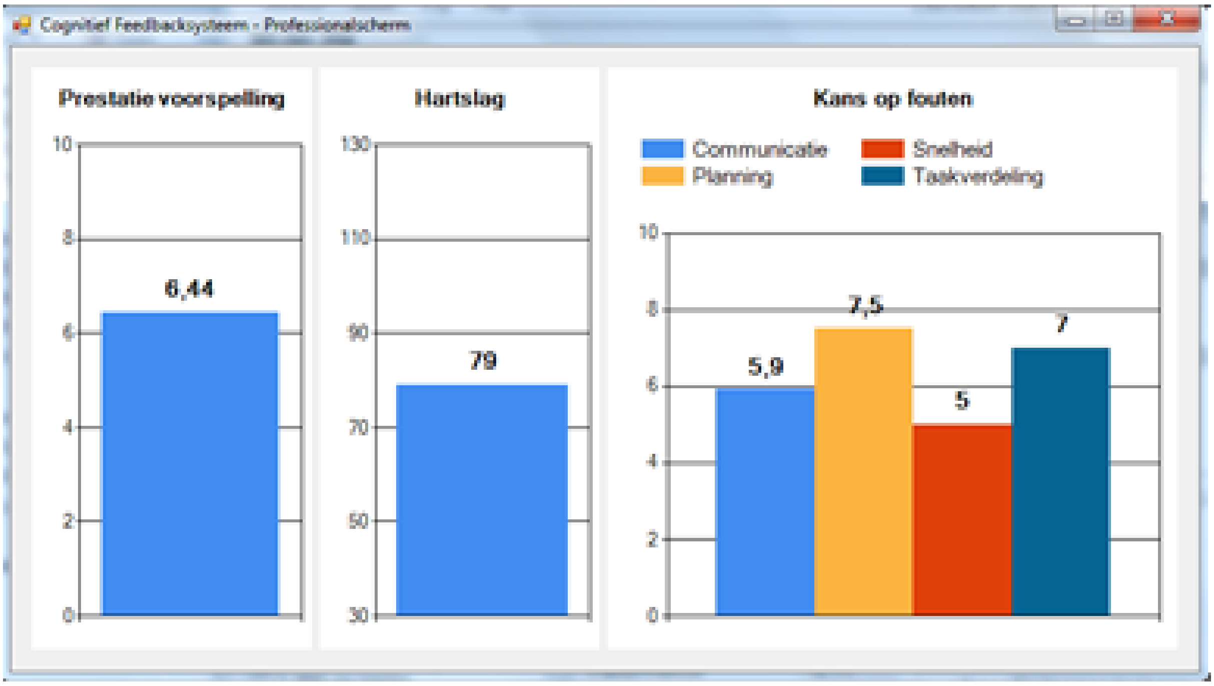


Figure 19: Verschillende soorten ondersteuning bij de taken.

Samengevat

Jouw taak is om branden af te handelen. Dat doe je door de volgende stappen uit te voeren:

- Selecteer de brand die je wil blussen.
- Stel de vragen die nodig zijn aan de hand van de beslisboom.
- Attendeer de ziekenboeg als dat nodig is.
- Selecteer de juiste actie.

Blijf goed op de beslisboom letten om uit te vinden wat de juiste actie is en of er nog aanvullende informatie is over de brand en onthoud dat je in bepaalde situaties hulp kunt vragen met een brand als je tijd tekort komt.

Denk eraan dat de deelnemer die de meeste levens redt een beloning van €20,- ontvangt. Aan het einde van het onderzoek hoor je hoe veel levens je gered of verloren hebt.

Voor je begint

Als dit nog niet gebeurd is mag je nu de hartslagmeter omdoen en zullen de proefleiders de nodige sensoren opplakken. Daarna mag je aan de tutorial en het oefenscenario beginnen. Het is belangrijk dat je tijdens het uitvoeren van de taak zo stil mogelijk zit om de meetapparatuur niet te verstoren of los te trekken.

Is alles duidelijk?

9 Appendix B - Dutch decision trees

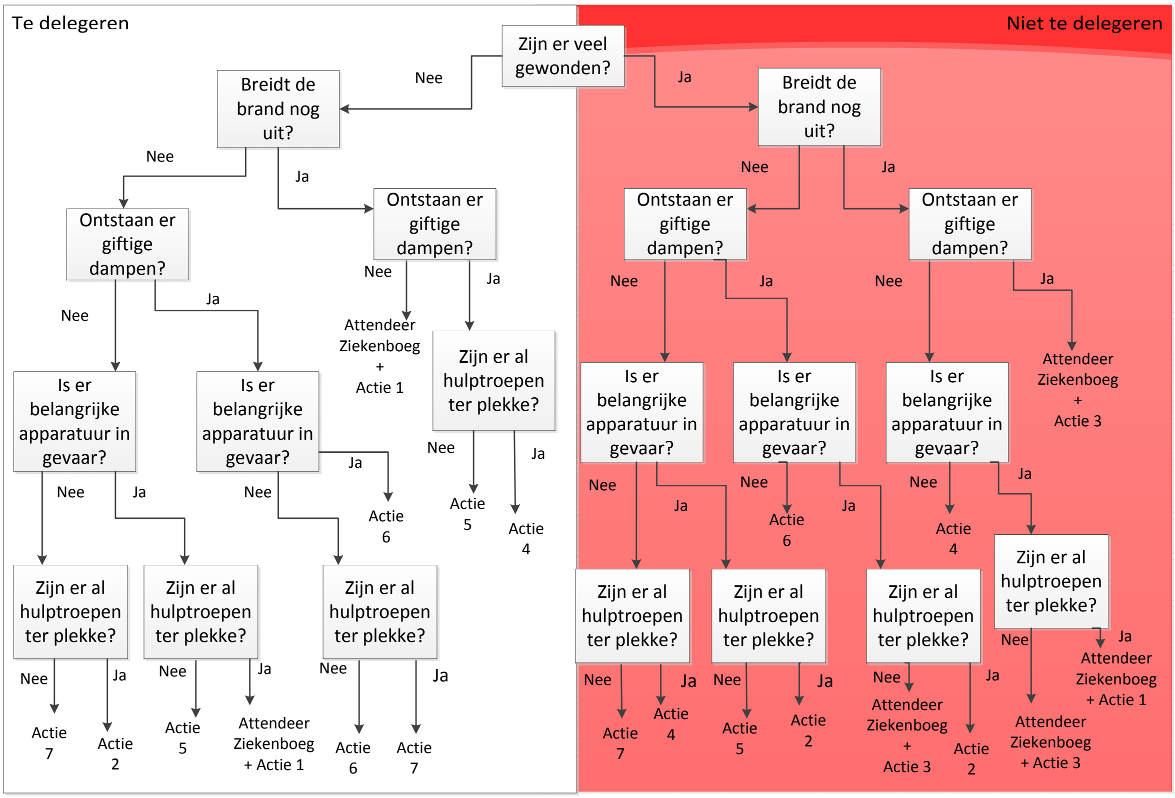


Figure 20: Decision tree for scenarios one and two.

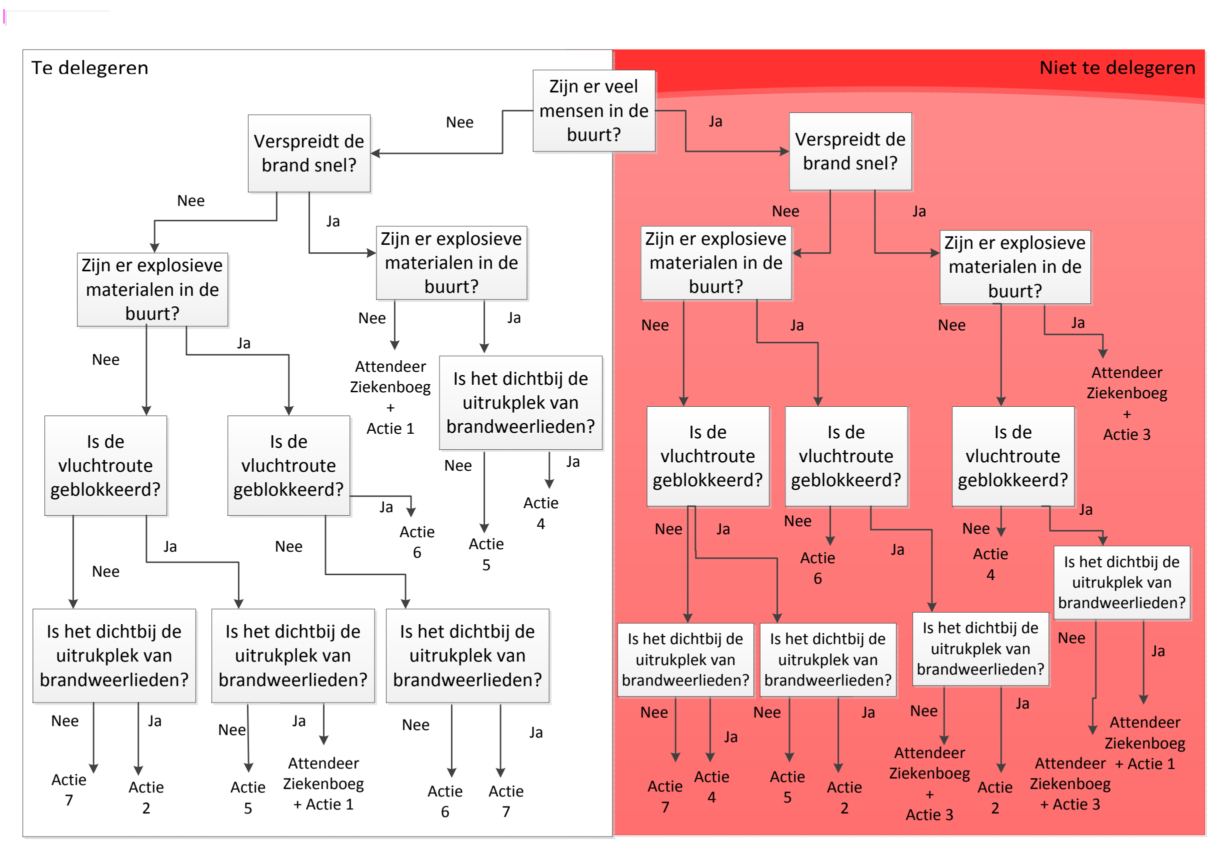


Figure 21: Decision tree for scenarios three and four.

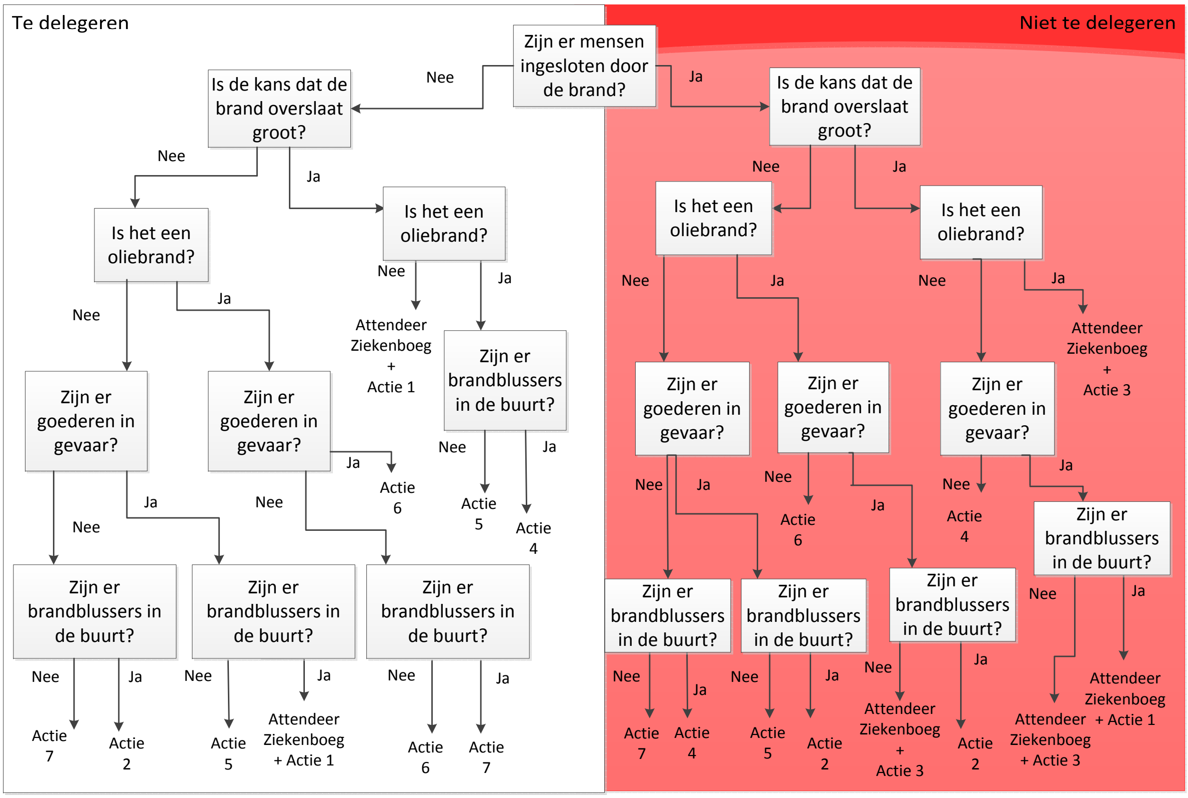


Figure 22: Decision tree for scenarios five and six.

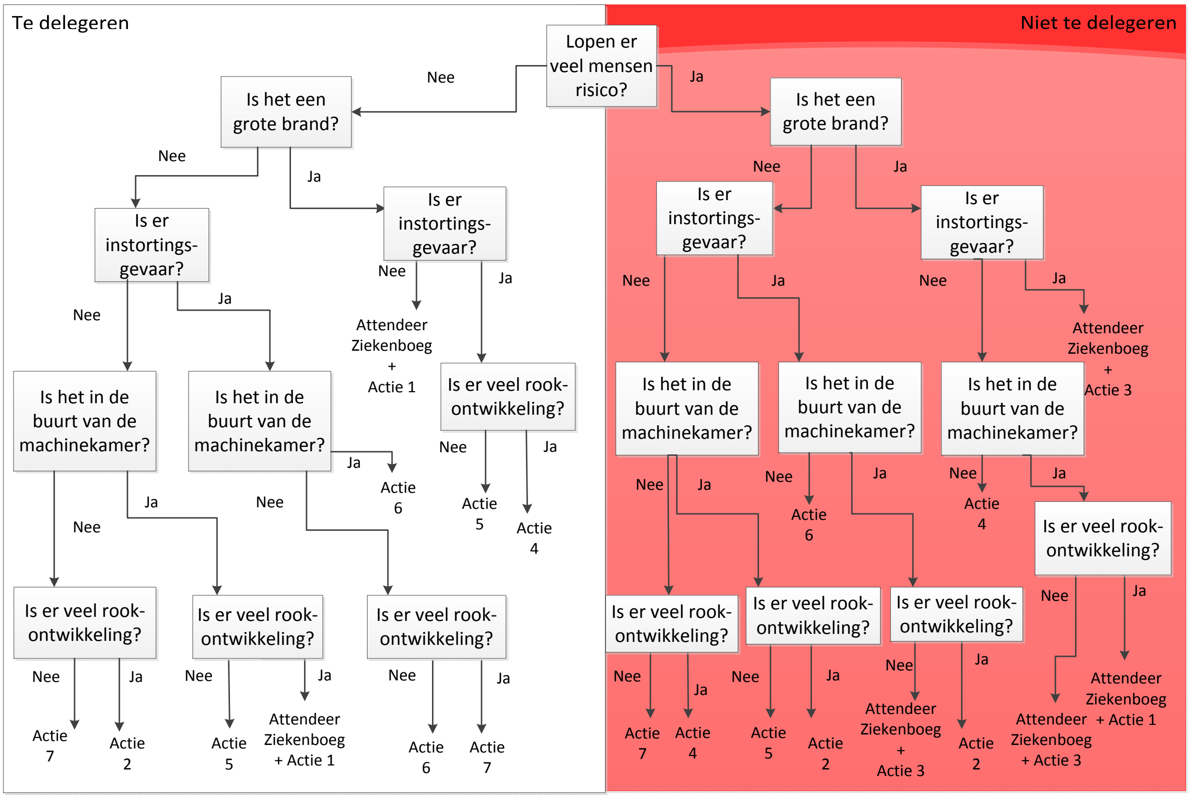


Figure 23: Decision tree for scenarios seven and eight.

10 Appendix C - Feedback conditions

typen feedback

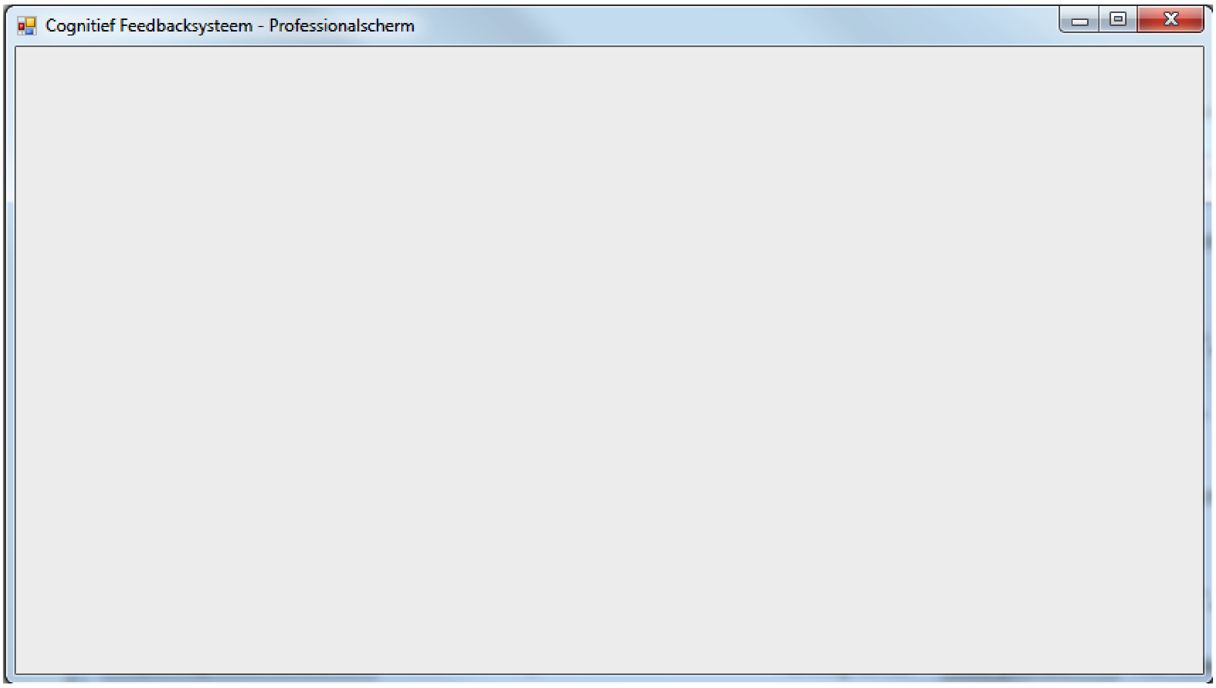


Figure 24: Condition 1, no feedback.

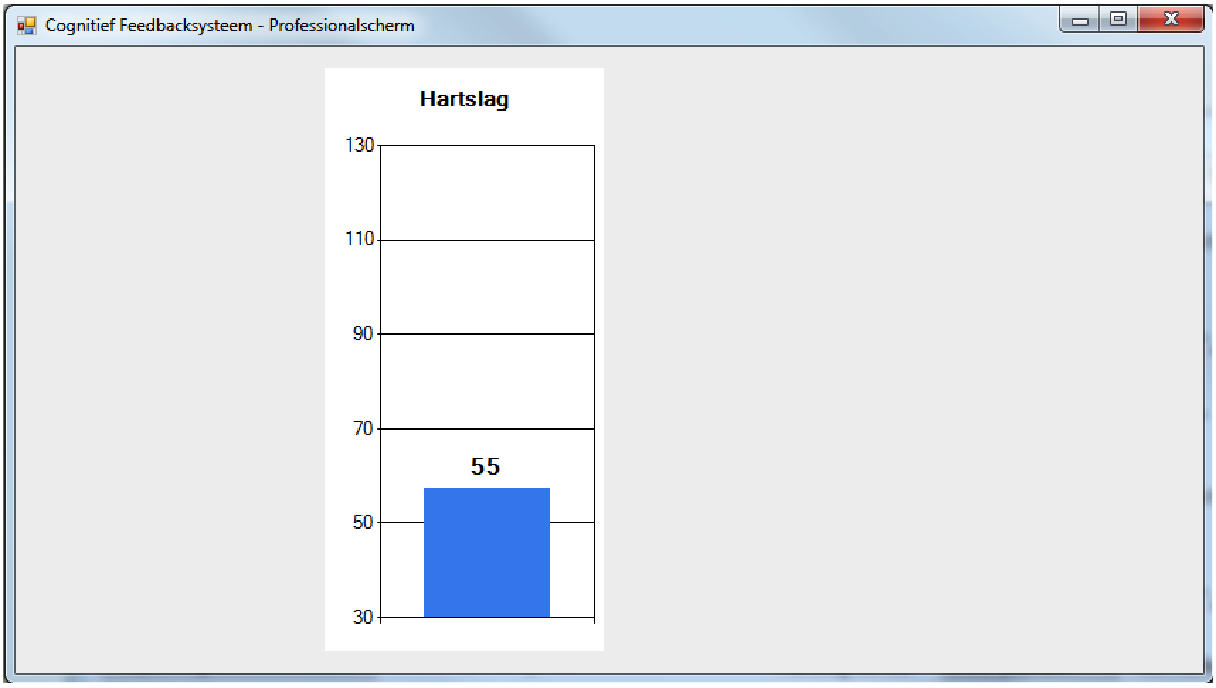


Figure 25: Condition 2, biofeedback.

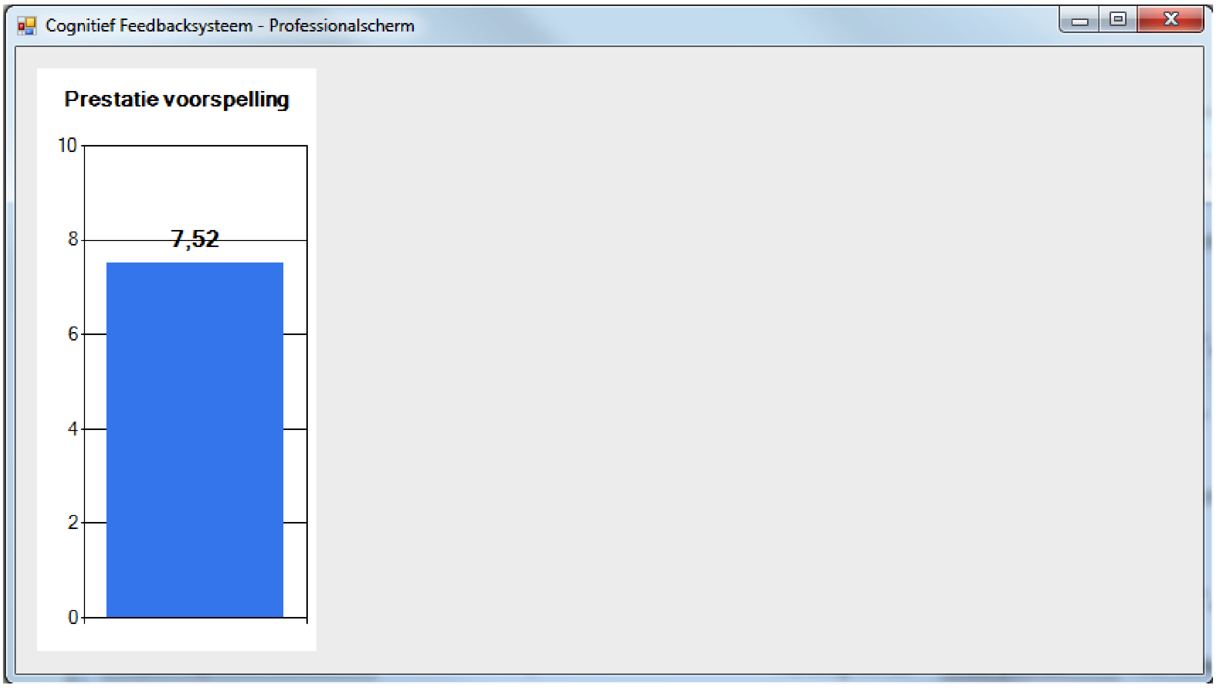


Figure 26: Condition 3, performance-prediction.

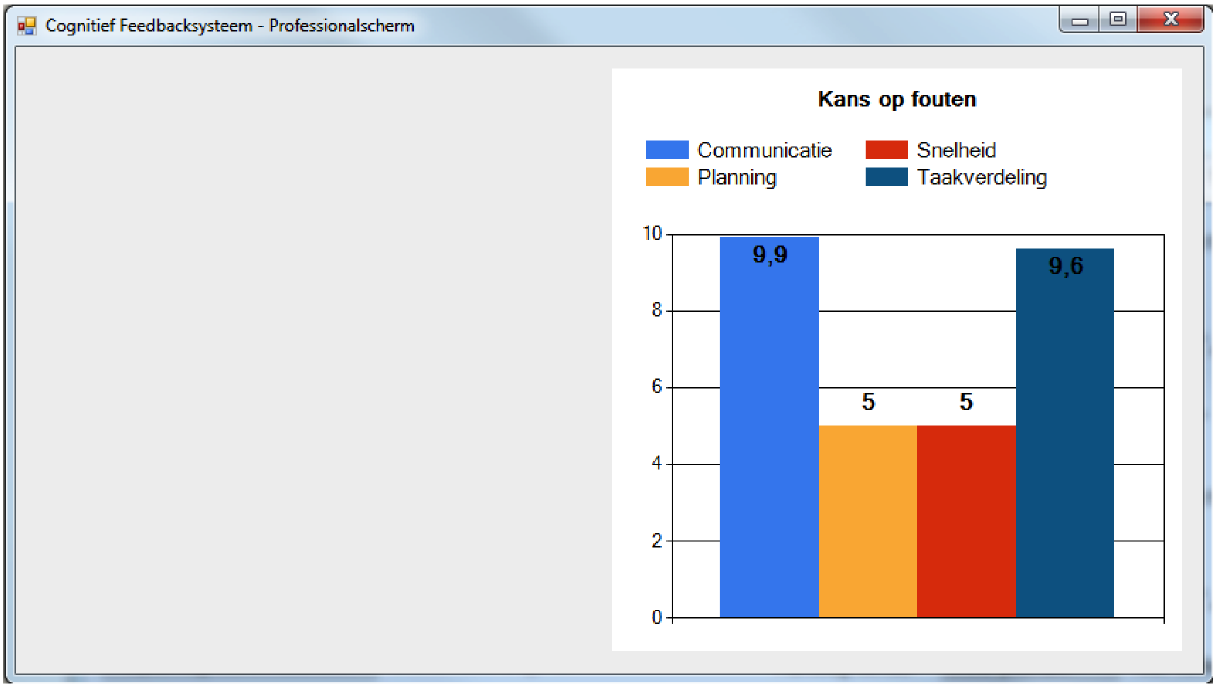


Figure 27: Condition 4, error-prediction.

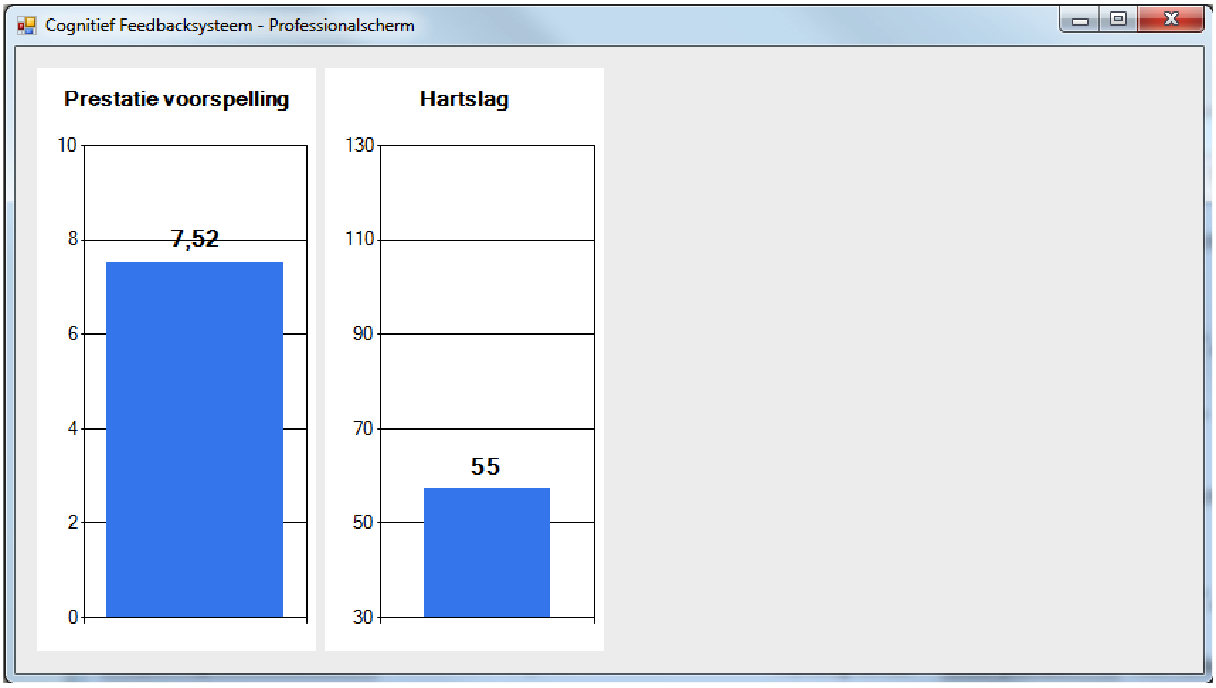


Figure 28: Condition 5, biofeedback and performance-prediction.

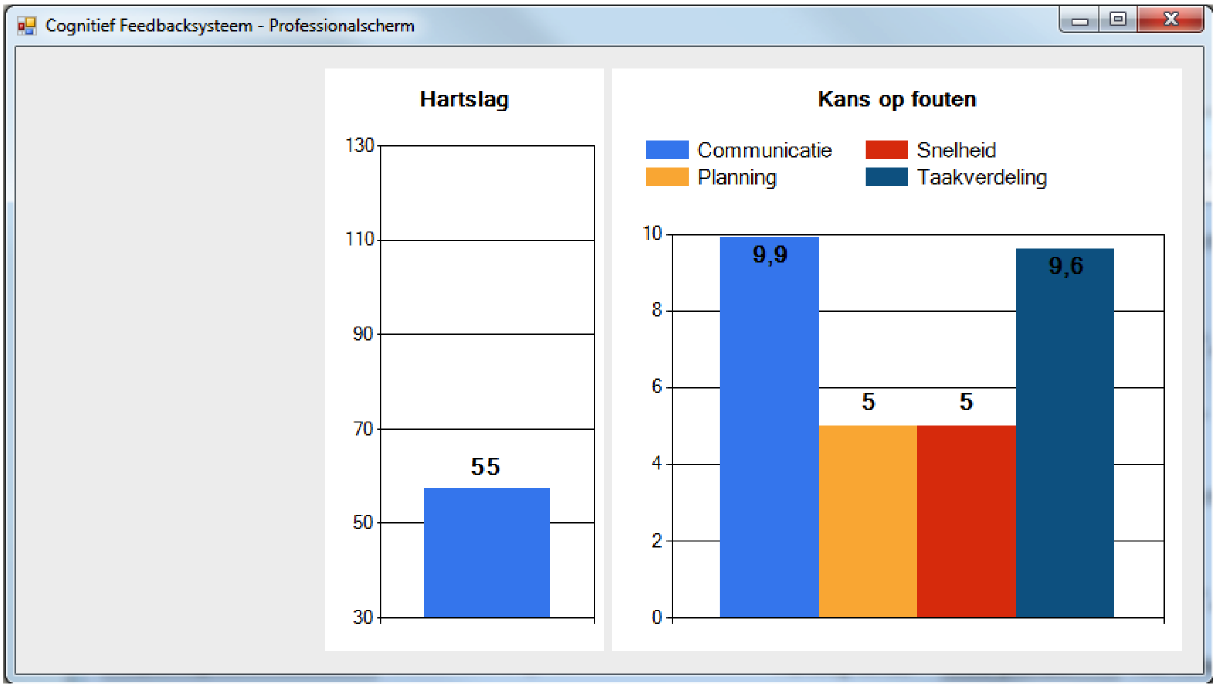


Figure 29: Condition 6, biofeedback and error-prediction.

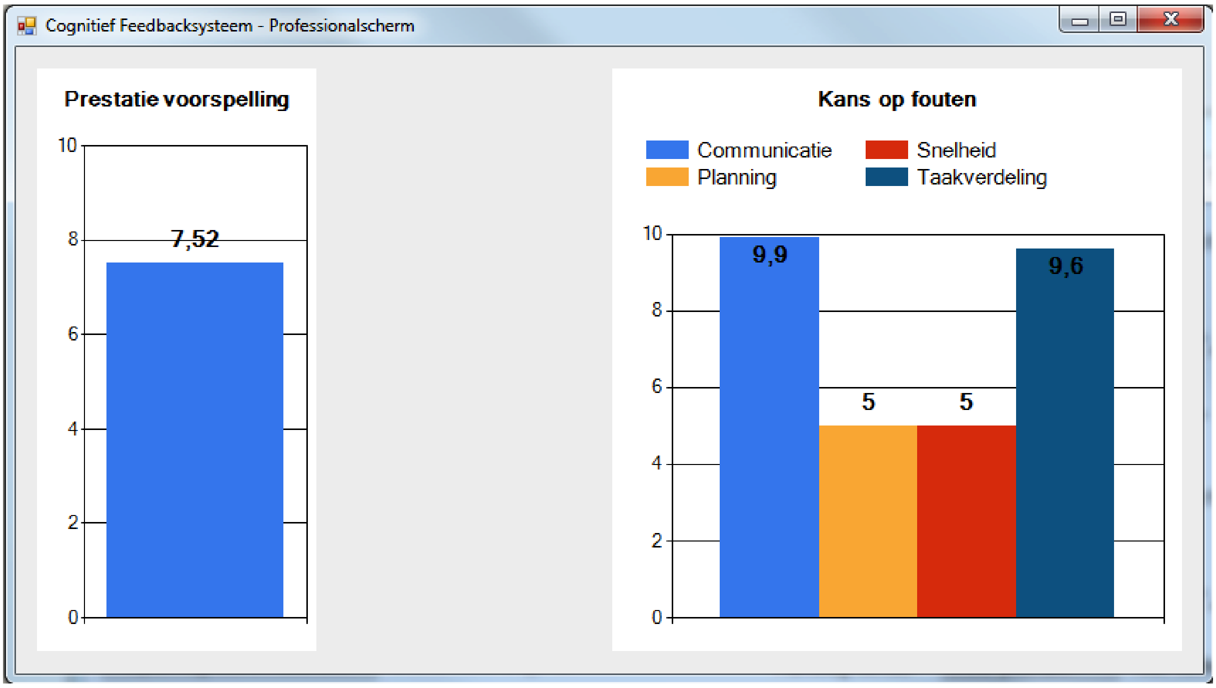


Figure 30: Condition 7, performance-prediction and error-prediction.

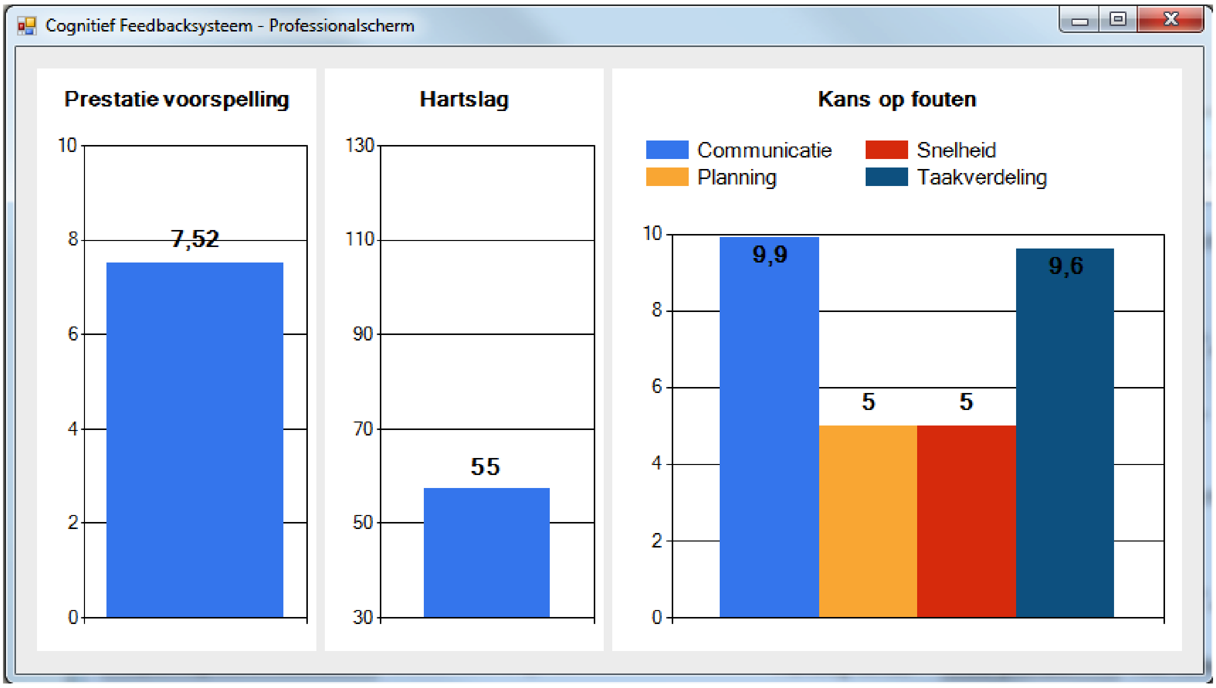


Figure 31: Condition 8, biofeedback, performance-prediction and error-prediction.

11 Appendix D - List of abbreviations

- CFSystem = Cognitive Feedback System
- COPE-model = COgnitive Performance and Error model
- HR = Heart Rate
- HRV = Heart Rate Variability
- LIP = Level of Information Processing
- SUS = System Usability Scale
- TO = Time Occupied
- TSS = Task Set Switches

12 Appendix E - List of translations for pictures and graphs

- Acties = actions
- Communicatie = communication
- Hartslag = heart rate
- Hier zijn 4 levens in gevaar. = There are 4 lives at stake.
- Hulp vragen = ask for help
- Kans op fouten = error-prediction
- Niet te delegeren = it is not possible to ask for help
- Planning = planning
- Prestatievoorspelling = performance-prediction
- Snelheid = speed
- Taakverdeling = task allocation
- Te delegeren = It is possible to ask for help
- Vragen = questions
- Ziekenboeg attenderen = notify sickbay