Computers in Human Behavior 99 (2019) 144–155

Contents lists available at ScienceDirect

Computers in Human Behavior

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

Determining the effect of stress on analytical skills performance in digital decision games towards an unobtrusive measure of experienced stress in gameplay scenarios

Johannes Steinrücke[a,*], Bernard P. Veldkamp[b], Ton de Jong[a]

[a] Department of Instructional Technology, University of Twente, P.O. Box 217, 7500AE Enschede, The Netherlands
[b] Department of Research Methodology, Measurement and Data Analysis, University of Twente, P.O. Box 217, 7500AE Enschede, The Netherlands

ARTICLE INFO

Keywords:
Interactive learning environments
Evaluation methodologies
Unobtrusive assessment
Crisis management
Serious games

ABSTRACT

This study aims to develop an unobtrusive measure for experienced stress in a digital serious gaming environment involving decision making in crisis management, using only in-game measures in a digital decision game called the Mayor Game. Research has shown that stress has an influence on a decision-maker's behavior, and also on the learning experience in training scenarios. Being able to assess unobtrusively the level of stress experienced would allow manipulation of the game so as to improve the learning experience. An experiment was conducted with two conditions, one paced and one non-paced. In the paced condition, participants were exposed to in-game changes that aimed to induce stress by creating information overload, uncertainty and time pressure. While pacing caused differences between the conditions with respect to in-game performance for analytical skills, several simple unobtrusive in-game measures were not consistent enough to serve as indicators for experienced stress. Further, physiological measurements of stress did not show significant differences between the conditions, indicating that the employed methods to induce stress did not work sufficiently. These results call for testing of more sophisticated methodologies to unobtrusively assess experienced stress in the given type of serious game.

1. Introduction

"Hasty, often unwise decisions" are induced by stress (Cooper, 2007, p. 21). This statement highlights the influence of stress on the quality of decision-making processes and was confirmed by later findings describing the effect of stress on decision making (Starcke & Brand, 2012). It is also related to the Yerkes-Dodson law (Yerkes & Dodson, 1908), which indicates that people perform worse under higher than optimal stress conditions. The Yerkes-Dodson law, describing a curvilinear (U-shaped) relationship between stress and performance, further implies that a slight level of experienced stress is nonetheless beneficial compared to experiencing no stress at all (Yerkes & Dodson, 1908). This relationship is of paramount importance when a high quality decision-making process is needed, such as decision making in the context of crisis management. In some crisis situations the decisions to be taken are similar to dilemmas: No matter what exact decision is made, the result is never perfect for all parties involved. In these situations, it is crucial to have a high quality decision-making process, thereby ensuring that decisions are not hasty and unwise, but well thought through, and that important factors are considered (Crichton, Flin, & Rattray, 2000).

One of the crucial aspects of crisis management is situation assessment, in which the available information and its meaning are examined and evaluated (Mezey, 2004). A thorough analysis of the crisis situation is also crucial for situational awareness (SA; Endsley, 2000), about which the author stated: "Most simply put, SA is knowing what is going on around you" (p. 2). Knowing that, we can reasonably conclude that situational awareness and situation assessment are key in decision making in crisis management (Mezey, 2004; van der Hulst, Muller, Buiel, van Gelooven, & Ruijsendaal, 2014; Veiligheidsregio Twente, 2016). Further looking at the role of stress, Mezey (2004) explained that stress can be produced in crisis situations, which would influence the crisis manager's performance. This stress could also influence the assessment of the crisis situation by decreasing the quality of the situation assessment and in the end reducing the quality of the decision-making process (Cooper, 2007). As Kowalski-Trakofler and Vaught (2003) described in the context of managing emergencies, stressed persons rely on different patterns of reasoning for coming to a conclusion than non-stressed persons, who more often rely on in-depth analysis of the situation. The authors further explained that decision

[*] Corresponding author.
E-mail address: j.steinrucke@utwente.nl (J. Steinrücke).

https://doi.org/10.1016/j.chb.2019.05.014
Received 8 November 2018; Received in revised form 8 May 2019; Accepted 10 May 2019
Available online 16 May 2019

makers often narrow their attention to focus on critical issues and elements when under stress, letting them ignore or miss other (additional) useful information about the situation. In the next section we explain how this relationship between stress and analytical skills is viewed and handled by the Veiligheidsregio Twente, a regional Dutch crisis organization, which also provides the practical context for this study.

1.1. Analytical skills and stress in the Veiligheidsregio Twente

The Veiligheidsregio Twente[1] (VRT; Twente Safety Region) is an organization consisting of people from the fire department, the police department, the medical emergency response team, and local government, responsible for coordination and response to crises in the Dutch region of Twente. The VRT usually involves two different types of crisis professionals in such crisis situations: First, there is the operational staff. These are, for example, the firefighters who go to the place of an incident and work on extinguishing the fire. Second, there is the strategic staff. They focus on the more administrative part of crisis management, which includes, for example, coordinating the different emergency services, specifying strategies and communicating facts within the organization and to the public. More complex crisis situations, where both types of crisis managers are involved, are only trained for about one afternoon once a month because of time and cost reasons. Consequently, training all important competencies (the VRT names 13 different competencies: *stress resistance, analyzing, empathy, leadership, teamwork, flexibility, decisiveness, persuasiveness, coordinating, communication, being advisory, political administrative insight, decision making*) and focusing on all of them is not feasible. Hence, the VRT has searched for an additional way to train their crisis managers on some of these competencies, for example, by a digital serious game that focuses on decision-making processes.

As laid out before, the two competencies of *Analyzing* (Analytical Skills) and *Stress Resistance* lie at the core of the decision-making process: The VRT (personal communication, April 16, 2018) describes Analyzing as the competency to discriminate between facts and assumptions, relevant and irrelevant information, to make connections based on available information, to gather and consult information as well as the ability to apply scenario-based thinking. Sarpong and Maclean (2011) define scenario-based thinking as the “use of scenarios to stimulate innovative solutions for a possible future context” (p. 1155). The VRT (personal communication, April 16, 2018) describes Stress Resistance*,* on a slightly different level. While the main point of being stress resistant is that the crisis manager does not get carried away by emotions, the description also implies that a stress-resistant crisis manager behaves similarly when under stress and when not under stress. This includes the requirement that the crisis manager maintains concentration and still puts things into perspective. According to the VRT, time pressure is not only among the most important stressors for crisis managers, but they also have to deal with time pressure during crisis situations, which does not allow them to discuss all matters in as much detail as possibly needed.

To ensure that the crisis managers of the VRT are capable of doing just that, more complex crisis situations thus need to be trained more often, for example, by means of a digital serious game. The experienced stress level of the crisis managers should be kept at medium intensity during training, which is where fastest learning can be expected according to Raudys and Justickis (2003), along with best performance as implied by the Yerkes-Dodson law. Knowing the player's current level of experienced stress would thus allow manipulation of the serious game to ensure that players experience the optimal stress level for learning, for example, by adjusting the pace of the game.

1.2. Decision games

Serious games are often categorized based on their application context (Susi, Johanneson, & Backlund, 2007) or on the game's genre (Connolly, Boyle, MacArthur, Hainey, & Boyle, 2012; Sniezek, Wilkins, & Wadlington, 2001). Since the current study was conducted within a crisis management and decision-making domain, the serious game to be used in this research would be categorized as a military or government game based on the scheme described by Susi et al. (2007). In both categories, managing crisis situations, making decisions and running scenarios repeatedly are among the most important game characteristics. However, this basic categorization does not spell out the specific nature of what a game fitting this research's context should look like.

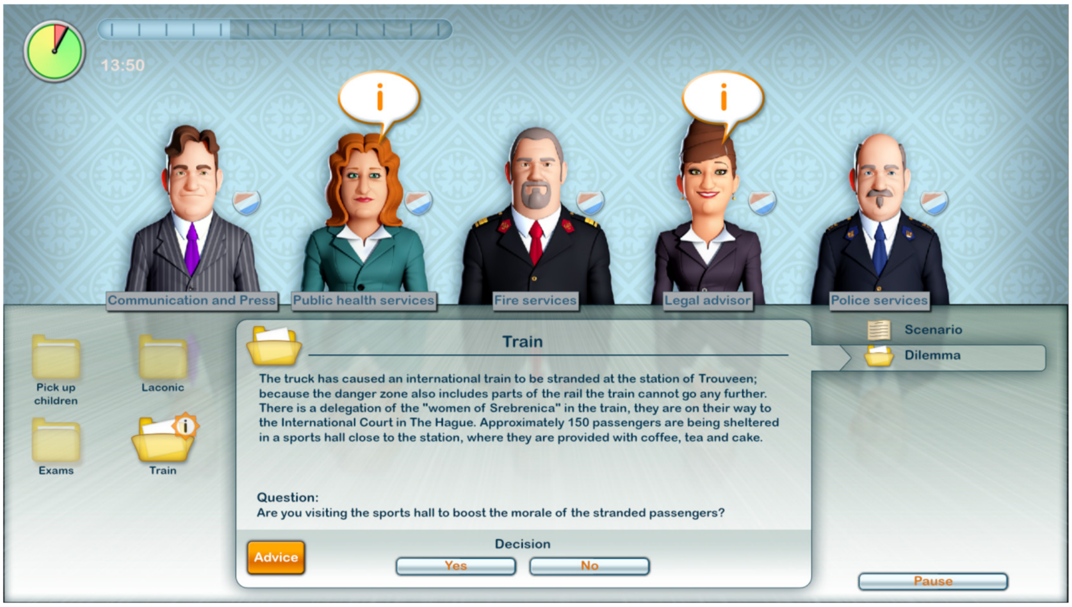
As Connolly et al. (2012) and Sniezek et al. (2001) described, many serious games and games for learning can be categorized as simulation games. In simulation games the goal is to model or mimic a realistic (crisis) scenario, for example, by using scenarios and/or dilemmas based on past, elapsed real-life crisis situations. In that sense, simulation games are well suited for training for decision making in crisis management. Just as the VRT already does in an analog manner, a serious game should mimic the situation for decision makers using realistic scenarios and dilemmas. As in real situations, the decision makers need to differentiate between pros and cons for a specific decision, in order to make a decision based on the available information. Sniezek et al. (2001) extended this description with the notion that in crisis (management) training contexts, “a simulation is realistic if it induces the same *psychological processes* in the training context that are experienced during an actual crisis” (p. 3), which remains one of the greatest challenges for such simulations to accomplish. Accordingly, we prefer to employ a serious game that mimics the to-be-trained-for situation in an authentic way, being aware that a simulation close to the real-life situation is difficult to implement.

What fits with the above categorizations, the context of our research, and also the structure of the VRT, is what Crichton et al. (2000) and Crichton and Flin (2001) described as tactical decision games, which are designed for practicing decision-making skills. Such games illustrate decision-making strategies, “boost expertise in decision making and judgement” (p. 260), but also allow practicing of situation awareness and stress management (Crichton & Flin, 2001). Further, the players must deal with the fact that there is no ‘correct’ decision to be made; the players are confronted with dilemmas. Players in such games also have only limited information available, in line with what occurs during actual crisis situations where decision makers must make a decision based on the available information, which sometimes is not much (Crichton et al., 2000). The *Mayor Game*, which is the digital serious gaming application used in this study, is focused on the process of making a decision, and functions similarly to what was just described in this section.

1.2.1. The Mayor Game

The Mayor Game is a digital web-based serious game used to train Dutch mayors for dealing with crisis situations. The Mayor Game provides training on various important competencies for decision making in crisis management, for example, *environmental consciousness*, *decisiveness*, *anticipation*, and *judgment.* The Mayor Game is played by many mayors all across the Netherlands, who in general value the game (Jong, 2017). In the game, players must handle a realistic crisis situation (called a scenario) by responding to a number of dilemmas. To help the players decide on what answer to give, a number of *advisors* offer additional information to the player, which the players can also mark as relevant. Next to additional information, each advisor also provides a clear recommendation on what to decide (*yes* or *no*). The game's focus does not lie on deciding correctly, given that there is no correct decision to make; instead, the Mayor Game focuses on the decision-making process, rather than on the actual decision. All dilemmas can be responded to in multiple ways, which will affect only the feedback the

[1] http://www.vrtwente.nl.



Train

The truck has caused an international train to be stranded at the station of Trouveen; because the danger zone also includes parts of the rail the train cannot go any further. There is a delegation of the "women of Srebrenica" in the train, they are on their way to the International Court in The Hague. Approximately 150 passengers are being sheltered in a sports hall close to the station, where they are provided with coffee, tea and cake.

Question:
Are you visiting the sports hall to boost the morale of the stranded passengers?

Fig. 1. Screenshot of the mayor game.

players receive after they have finished the scenario. See Fig. 1 for a screenshot taken in the Mayor Game, showing one dilemma. The feedback is about how the players handled the situation, what the players took into account and how the players scored on, for example, scales for different leadership styles (T-Xchange, 2018; van de Ven, Stubbé, & Hrehovcsik, 2014). Tweaked versions of one game scenario in the Mayor Game, as described in the next section, were used in the experiment described in this study.

1.3. An unobtrusive measure of stress in decision games

Unobtrusively assessing players’ stress levels in serious games has various advantages over more obtrusive in-game assessment methods such as questionnaires or external, out-of-game measures such as wearable sensors to measure physiological indicators of stress. For example, Shute (2011) stated that keeping the players in a so-called flow state by not interrupting or disturbing the gameplay to assess the players is beneficial to the learning process. The players are allowed to simply play the game, while unobtrusive in-game assessment methods can be used to induce adaptivity by changing the game depending on the outcomes of that assessment (Bellotti, Kapralos, Lee, Moreno-Ger, & Berta, 2013), which in turn increases the efficiency of the learning process for which the serious game was originally designed (Lopes & Bidarra, 2011). Further, the authors stated that adaptivity is supposed to increase how the game appeals to the players, which Cocea and Weibelzahl (2009) framed as follows: The more appealing the game, the more motivated the players, which in turn leads to increased learning.

To assess the players unobtrusively, the assessment must be woven into the natural gameplay (Shute & Kim, 2014). Consequently, the interactions of the players with the game need to be used, in such a way that conclusions can be drawn about the players’ experienced stress level. For example, we could just look at the percentage of correct decisions players have made up to a certain point. However, when a serious game is designed to improve the quality of processes and not the quality of their product, we cannot look only at the percentage of correct decisions or actions. Instead, we have to take into account other interactions from the player with the game that the player took in working towards said product.

Looking back at what has been described thus far, the current study aims to find an unobtrusive measure for experienced stress in crisis management-related decision making using different in-game measures in a digital games-based learning environment. The study is conducted in the context of the Veiligheidsregio Twente and the Mayor Game, a decision-based digital serious game.

1.4. Research questions

Finding an unobtrusive measure for a player's experienced level of stress is only useful when stress indeed has an influence on in-game performance. Therefore, we investigate on the effect of induced stress on in-game performance first. Given that it is crucial for decision makers in crisis management to maintain the quality of their decision-making process while under stress, in the case of the VRT, the ability to analyze information, in particular, should not suffer under greater experienced stress. This gives rise to the research question: *Does a change in experienced stress influence the players' in-game performance for analyzing?* Based on the definitions stated by the VRT, it is expected that persons experiencing too much stress will score lower on analyzing, which is in line with Yerkes and Dodson's (1908) theory that a higher than optimal level of experienced stress causes worse performance. Answering this research question will help to identify effects of induced stress on players' performance for analyzing in decision games that provide training on crisis management skills, thereby highlighting the importance of stress in the crisis managers' everyday work.

Next, in line with the Yerkes-Dodson law, which describes a curvilinear (U-shaped) relationship between the players' stress levels and their performance (Cohen, 2011; Yerkes & Dodson, 1908), stress levels that differ from what is optimal are expected to have a negative influence on a person's response behavior (e.g., Cheng, 2018; Sniezek et al., 2001). Assuming that the level of experienced stress indeed has an influence on the players' in-game performance, the second research question of this study is: *Can we employ in-game behavior to make predictions about the measured level of stress of a player?* Answering this research question serves the general aim of this study to find an unobtrusive measure for experienced stress in crisis management-related decision making in a digital games-based learning environment. This unobtrusive measure could be used as input to statistical models that can evaluate the experienced level of stress to make the game adaptive. Thereby, the unobtrusive measure allows to guide the game to be more effective and efficient: After a calibration scenario, to obtain base levels of the player's stress level, the game can adjust the following scenarios based on a comparison of the current stress level of the player as measured by the unobtrusive in-game measure and the earlier set baseline.

In the current study, participants will be placed in two conditions, one control and one experimental condition, to gain insight into how stress affects their in-game behavior and their performance for analytical skills. Gameplay log data, questionnaire data, and physiological sensor data will be used to address the research questions.

2. Method

2.1. Participants

In total, 82 participants took part in this study, 43 in the control condition and 39 in the experimental condition. Participants were randomly assigned to condition, but due to short-term cancellations the group sizes ended up different. Two participants in the experimental condition did not complete the experiment and were thus removed from the dataset. The remaining 80 participants (43 in the control condition and 37 in the experimental condition) came from two different programs of study at the University of Twente, Psychology (59; 73.75%) and Communication Science (21; 26.25%). The majority of the participants were female (53; 66.25%). Participants had various nationalities, the two most common being German (58; 72.5%) and Dutch (15; 18.75%). The remaining seven participants were scattered among five different nationalities, so they were grouped under Other (7; 8.75%). The great majority of participants were between 18 and 24 years old (75; 93.75%), where only five participants were 24 or older (6.25%).

2.2. Game scenario

Participants played one scenario in the Mayor Game: A tanker truck transporting highly flammable and toxic chemicals was involved in an accident and threatens to explode. The scenario consisted of eight dilemmas; participants decided about each dilemma with either *Yes* or *No*. In the dilemmas the players had to make decisions about how to handle different situations. For example, the first dilemma, named “Pick up Children”, was to decide whether parents should be allowed to pick up their children from a school within the evacuation radius. The other dilemmas were called “Laconic”, “Exams”, “Train”, “Tight Curve”, “Laboratory”, “Youth News” and “Explosion”. “Laconic” is a dilemma about how to deal with residents in the wider area who do not listen to the instructions made by the police. The third dilemma, “Exams”, asked whether final exams taken in a school lying in the evacuation area should be stopped. The fourth dilemma, “Train”, confronted the players with the question whether a shelter hall where stranded passengers are held should be visited by the mayor of the fictional town in which the truck accident took place. Dilemma five, “Tight Curve”, challenged the players to deal with a statement that was published by a third party, trying to explain how the accident could happen in the first place. Dilemma six, “Laboratory”, confronted the players with deciding on whether a laboratory experimenting with viruses should receive additional protection against the smoke and dust coming from the burning truck. “Youth News” was a dilemma asking the players to handle the press, which focuses on explaining the situation to a younger audience. In the eighth and last dilemma, “Explosion”, the tanker truck exploded, so the players faced the challenge of deciding how the death of three firefighters should be communicated to their respective families.

Five different advisors (police services, fire service, public health services, legal advisor, communication and press) could help the players with making their decisions by providing additional information on each dilemma. Each advisor provided one information item per dilemma, adding up to five information items available to the players per dilemma and consequently 40 in total. Each could be read multiple times before responding to the dilemma. Although the eight dilemmas were offered to the players in a pre-specified order, the players were not forced to play them in this particular order. Participants were asked to mark the information corresponding with the decision that they made, to indicate which information implied that same decision. As an example, we provide a detailed description of one dilemma and the five information items offered for this dilemma in Appendix A.

2.2.1. Conditions

Participants in this study were assigned to one of two conditions. In the control condition participants were given 15 min to complete the scenario. Dilemmas and information provided by the five advisors appeared at fixed time intervals, which were the same in all dilemmas. For example, when a participant opened the first dilemma, information items became available every two to 4 s. Additionally, a recommendation on whether an advisor would vote *yes* or *no* appeared every 20 s, with two being available from the start. That resulted in a well-suited game pace, where a new action became available whenever the players finished thoroughly reading information items or descriptions of the situation.

A heightened stress level can be created by inducing information overload and time pressure (see Cheng, 2018; Sniezek et al., 2001), which is in line with the in section 1.1 described information provided by the VRT. Thus, participants in the experimental condition were given 12 min to complete the scenario, 3 min less than in the control condition, but still enough time to complete the scenario without skipping information. To further increase the experienced stress level, information given by the advisors appeared faster. Furthermore, the preferred decisions of the advisors were available much later in the experimental condition, so that uncertainty was created if the participant had not yet made a choice. Uncertainty is also used to increase the experienced stress level (Sniezek et al., 2001). Cheng (2018) also stated that a (large) alarm clock, visible and/or audible to the participant, might also increase time pressure. In accordance with this suggestion, a large alarm clock was presented on a screen on location, along with a verbal statement of the time left made at 2-min intervals.

Along with time pressure, an equivalent to peer pressure was induced. Kim, Bang, and Kim (2004) used simple statements such as: “The participants before you all did well and finished relatively fast” to increase the stress level of the participant. Consequently, a similar statement was made in the experimental condition, before the participants started the gameplay scenario. The content of the scenario was not altered in comparison to the control condition, so that possible differences in gameplay behavior and stress levels between the conditions were fully explainable by the artificially induced time and peer pressure. By inducing time pressure, changes in decision-making processes would be expected to occur (Maule, Hockey, & Bdzola, 2000).

2.3. Measures

2.3.1. Analytical performance measure

According to the definition provided earlier for analyzing, interpreting available information correctly given the situational context is of utmost importance. In the context of the Mayor Game, available information items point towards either *Yes* or *No*. In some cases the information items are more vague and in some cases less vague, but they are always relevant to the situation. This leaves the players to interpret the information, and later to indicate which information corresponded with the decision they made. In doing so, the players indicated that the available information was correctly understood and the main message of this information item was grasped. Consequently, if players read all available information, they had to mark those information items as corresponding that implied the decision they made for that specific dilemma. When an information item did not support the player's decision, the player was expected not to indicate this item as corresponding. In line with the definitions of crisis management skills, any information that was not read was considered to be wrongly interpreted, because requesting information to reduce uncertainty is also part of these crisis management skills (Mezey, 2004; Veiligheidsregio Twente, 2016).

Players received one point for each information item correctly marked as corresponding and one point for each information item correctly non-marked, adding up to a maximum score of five points per dilemma equal to the amount of available information. Again, when information was not read it was considered as wrongly interpreted. Some of the information were vaguer than other, which besides simply computing scores also must be taken into account. For this reason, we used an Item-Response Theory (IRT) model, the Rasch model (Baker, 2001; Rasch, 1960), thereby bringing the person's ability and the item difficulty (here vagueness of information) onto the same scale. In this way we can estimate the person-level ability parameter for each person for each dilemma, taking into account the vagueness of the information items.

2.3.2. Behavior variability as unobtrusive stress measure

Since the same competency, analyzing, was measured in all dilemmas, under normal conditions the observed ability of a person is expected to remain stable throughout a single gameplay scenario in the Mayor Game. Participants under stress are expected to be less accurate in interpreting the information items and more inconsistent in playing the game, possibly leading to larger variability in terms of overall performance and gameplay behavior. Therefore, we expect larger variability throughout the scenario to be an indicator for induced stress. To assess the variability of the analytical ability parameter, the variance between the per dilemma parameter estimates can be computed per person. The resulting within-person variances are consequently considered to be an indicator of the variability of the ability parameter. We hypothesized that a more stressed person shows more inconsistent behavior; hence, we expect the within-person variances to be higher in the experimental condition than in the control condition. In line with this expectation, the within-person variances of three in-game measures were considered: of the average number of information items read, of the average time an information item was kept open relative to the total playtime, and of the performance on analytical skills throughout the gameplay.

2.4. Validation of the unobtrusive stress measure

2.4.1. Sensor data

To obtrusively measure experienced stress, sensor data was collect to validate possible in-game measures, so that in an actual practical application the sensors would not be needed to infer the player's level of experienced stress. To objectively measure experienced stress, we used the Shimmer GSR + sensors, which capture heart rate data and also skin conductance data. Both have shown to be valid predictors of physiological arousal (e.g., Mandryk, Inkpen, & Calvert, 2006; Shi et al., 2010; Yannakakis & Hallam, 2008), which in turn can serve as an indicator for experienced stress (Holmgård, Yannakakis, Martínez, Karstoft, & Andersen, 2015).

More specifically, Holmgård et al. (2015) found different measures extracted from sensor data that were significantly correlated to experienced stress: maximum heart rate, heart rate at last measurement point, heart rate standard deviation, heart rate range (the difference between maximum and minimum heart rate), skin conductance range and skin conductance at the first measurement point. The authors also found physiological measures that were statistically significant only with $\alpha = 0.1$. These measures were the average skin conductance throughout gameplay, maximum and minimum skin conductance, standard deviation of skin conductance and shift in heart rate throughout the game (the difference between first and last measurement of heart rate), all of which we will consider in our analysis. Taking into account that the Yerkes-Dodson law also suggests a U-shaped relationship between analytical ability (respectively the within-person variances of analytical ability) and (physiological) arousal, curvilinear as well as linear relationships between gameplay performance and sensor data were tested.

2.4.2. Self-report questionnaire

A self-report questionnaire was administered right after the participants finished the gameplay scenario. Here we followed the same idea as with the collected sensor data, using two brief (perceived) arousal scales as indicators for the stress level that was experienced. The *Perceived Arousal Scale* (Anderson, 1995; Anderson, Deuser, & DeNeve, 1995) and the *Affect Grid* (Russell, Weiss, & Mendelsohn, 1989) were used. Besides demographic characteristics and the arousal scales, participants were also asked to describe how they felt about the pace of the game in an open question format. The Perceived Arousal Scale is a 24-item questionnaire, where participants must indicate how much they felt a specific emotion or feeling using a five-point Likert scale, with one being *not at all* and five meaning *extremely*. The Affect Grid is a single item questionnaire measuring two dimensions, *arousal* and *pleasure*. The participants here must indicate how aroused they were and how much pleasure they felt during the gameplay session, this time on a nine-point Likert scale. The pleasure dimension was disregarded in this study.

2.5. Procedure

Participants were first familiarized with the game they were about to play. Important gameplay mechanics and features of the game were presented. The sensors were put on and immediately turned on, right after the participants were familiarized with the game. Next, the participants played the gameplay scenario. Once the participants were finished playing the gameplay scenario, they started completing the self-report questionnaire, while still wearing the sensors. Recording of the sensor data was stopped after the participants completed the self-report questionnaire. The researchers were present in the room during the entire time of the study to ensure that the participants followed the instructions; participants were able to ask for assistance in case of technical difficulties. Ethical approval for this research was provided by the ethics committee of the University of Twente in February 2018. All participants provided active consent before starting the study.

Data were collected using three main types of source. First, gameplay data from the Mayor Game were gathered. Second, sensor data about the heart rate and the galvanic skin response of participants was gathered using the *Shimmer GSR +* sensors. Third, a self-report questionnaire was employed. The sensor data and the self-report data were collected for validation purposes.

2.6. Analysis

Data were analyzed with respect to the two formulated research questions. First, we compared the analytical skills performance between groups, on a per dilemma basis. This was done based on simple sum scores, as derived from the scheme described in Appendix B and as explained in Section 2.3.1. To correct for differences in the vagueness of the available information, a one-parameter logistic (Rasch) model was employed. Given that only five observations per dilemma per person were available, we employed a Bayesian application of the Rasch model, making sure that the results are meaningful even with few(er) data points (McNeish, 2016; van de Schoot, Broere, Perryck, Zondervan-Zwijnenburg, & van Loey, 2015). Here we made use of an *Importance Sampling* (Wasserman, 2004, pp. 403–433) based procedure, as described in Appendix C. These analyses serve to address our first research question.

The second research question was answered using regression analyses. We started by comparing in-game performance and in-game behaviors between groups. Following that, simple regression models employing only one in-game measure and one physiological measure were used to gain insight into possible linear or curvilinear relationships between in-game data and physiological arousal data, that can serve as an indicator for experienced stress. We ran the same analyses for the in-game measures and the self-report data, where we also looked into the reliability of said self-reports. Lastly, we checked relationships between sensor and self-report data.

3. Results

3.1. The effect of stress level on 'analyzing'

All participants were scored for analyzing according to the scheme[2] described in Section 2.3.1. The lowest observed per dilemma score was zero, and the highest observed per dilemma score was five. The lowest observed total score was 14 and the highest was 40, meaning that at least one participant interpreted all information correctly. Table 1 displays the mean scores and standard deviations (SD) for all dilemmas and the total score, per condition. Levene's test for equality of variances between conditions for the total score gave a significant result, with $F = 5.903$ and $p = 0.017$, meaning the variances are not equal between groups, which indicated that the Welch-Satterthwaite adjustment method should be used on the degrees of freedom. Running a one-tailed *t*-test on the total score, using this adjustment method, gave a significant difference in total scores between conditions ($t = 2.777$; $df = 64.638$; $p = 0.0035$), indicating that the average score in the experimental condition was lower compared to the normal condition. Although this difference might seem interesting at first, it must be noted with caution: Because participants in the experimental condition were forced to work faster than participants in the control condition, some were not able to complete the gameplay scenario. Consequently, analyses should be carried out that take into account that some dilemmas might not have been accessed.

Following the order used in Table 1, from top to bottom, the first four dilemmas were completed by all 80 participants. The fifth dilemma was completed by 79 (43 control, 36 experimental) participants. The sixth, seventh and eighth dilemma were completed by 76 (42 control, 34 experimental), 73 (41 control, 32 experimental), and 68 (38 control, 30 experimental) participants respectively. Of the 12 participants who did not complete the eighth and final dilemma, five came from the control and seven from the experimental condition. However, not all of the 68 participants completing all eight dilemmas actually read all available information items. As already described, what information was not accessed could be determined from the log files, and it was considered as wrongly interpreted.

Excluding those participants who did not finish the scenario (all dilemmas) from the analyses gives a more accurate understanding. As described, in the control condition, 38 participants out of 43 total completed the scenario. In the experimental condition, 30 out of 37 participants completed the scenario. A one-tailed *t*-test on the total scores of these 68 participants, again using the Welch-Satterthwaite adjustment method on the degrees of freedom ($F = 5.875$; $p = 0.018$), showed that participants scored significantly higher in the control condition than in the experimental condition ($t = 2.178$; $df = 50.455$; $p = 0.017$). The per dilemma differences in scores between conditions are visualized in Fig. 2.

At the dilemma level, significant differences between conditions in the second, third, and last dilemma were found. These findings are partly in line with the feedback given by participants. In the experimental condition, participants reported that especially for the last dilemma, when they realized they would not finish the scenario in time at their current speed of working, they skimmed faster through the available information, not having as much time to think about what the information implied. Consequently, these participants made more mistakes in indicating which information also implied their decisions. However, this does not explain the significant differences in dilemmas two and three.

Table 1
Mean sum scores and standard deviations of dilemma and total scores per condition; One-tailed *t*-test statistic with fitting adjustment method for degrees of freedom.

Dilemma	Control	Experimental	*t*-test statistic
	Mean (SD)	Mean (SD)	t (df; p)
1. Pick up children	3.74 (1.432)	3.65 (1.602)	0.282 (78; 0.3895)
2. Laconic	4.14 (1.082)	3.54 (1.095)	2.455 (78; 0.008)**
3. Exams	4.60 (0.695)	4.22 (0.947)	2.063 (65.115; 0.0215)**
4. Train	4.26 (0.902)	4.00 (1.000)	1.203 (78; 0.1165)
5. Tight curve	4.07 (0.936)	3.73 (1.217)	1.411 (67.084; 0.0855)*
6. Laboratory	3.72 (1.368)	3.49 (1.539)	0.721 (78; 0.2365)
7. Youth news	3.86 (1.246)	3.32 (1.796)	1.527 (62.755; 0.066)*
8. Explosion	3.26 (1.720)	2.32 (1.857)	2.328 (78; 0.011)**
Total Score	31.65 (4.956)	27.27 (6.332)	2.628 (67.811; 0.0055)**

Note. $^{**}p < 0.05$, $^{*}p < 0.1$.

As already discussed, IRT analyses can be employed to account for differences in difficulty/vagueness of available information. By taking into account the complexities of the different information items, the Rasch model gives more accurate estimates of how well players performed and is more informative when it comes to drawing conclusions about the current ability levels of the players compared to score-based measures that do not take into account item difficulty. The difficulty of the information items was estimated using a Rasch model (e.g. Hambleton & Swaminathan, 1985), prior to feeding the values into the analyses presented in this study. All analyses were carried out using the data of the 68 participants who finished all dilemmas.

On average, participants performed better in the control condition than in the experimental condition. *T*-tests comparing the per dilemma person ability parameters between conditions supported the earlier findings based on per dilemma scores, thereby confirming that the experimental manipulation indeed led to a difference in in-game performance for analyzing between the two conditions. Person ability parameters, estimated using the Rasch model, are visualized in Fig. 3, again per dilemma and per condition. The IRT analyses confirmed the earlier score-based findings. By correcting for the item difficulty and bringing person ability and item difficulty onto the same scale, the IRT analyses introduced a "cleaned-up" ability estimate, that we can use to assess the variability of the analysis scores throughout the scenario. A 2-parameter logistic model (2PLM) or a 3-parameter logistic model (3PLM) might lead to even more accurate estimates of the person ability parameter. However, because of the relatively small size of the data set for estimating the item parameters, we chose to report the Rasch model analyses. Furthermore, an analysis conducted using a 2PLM resulted in issues in terms of model fit, while it also did not lead to meaningfully different results.

3.2. Unobtrusively assessing 'stress resistance'

3.2.1. Differences in (variability of) gameplay behavior between conditions

As a measure of experienced stress, higher within-person variances in the experimental than in the control condition were hypothesized. The within-person variances in the person ability parameter per condition are visualized in Fig. 4. To test for significance, linear regression or *t*-tests are not applicable, since the variance is not normally or *t*-distributed. In Bayesian solutions, often inverse gamma distributions are used to sample from the variance of a normal distribution (Lynch, 2007). Accordingly, a generalized linear model using an inverse gamma link, from here on referred to as inverse gamma regression, was used to test for between-condition differences in the within-person variances. Condition was found to have no significant effect on within-person variances ($\beta_1 = -0.4134$; $t = -1.111$; $df = 65$; $p = 0.271$). Hence, in the next section we used the collected sensor data to test for possible

[2] Provided in detail in Appendix B.

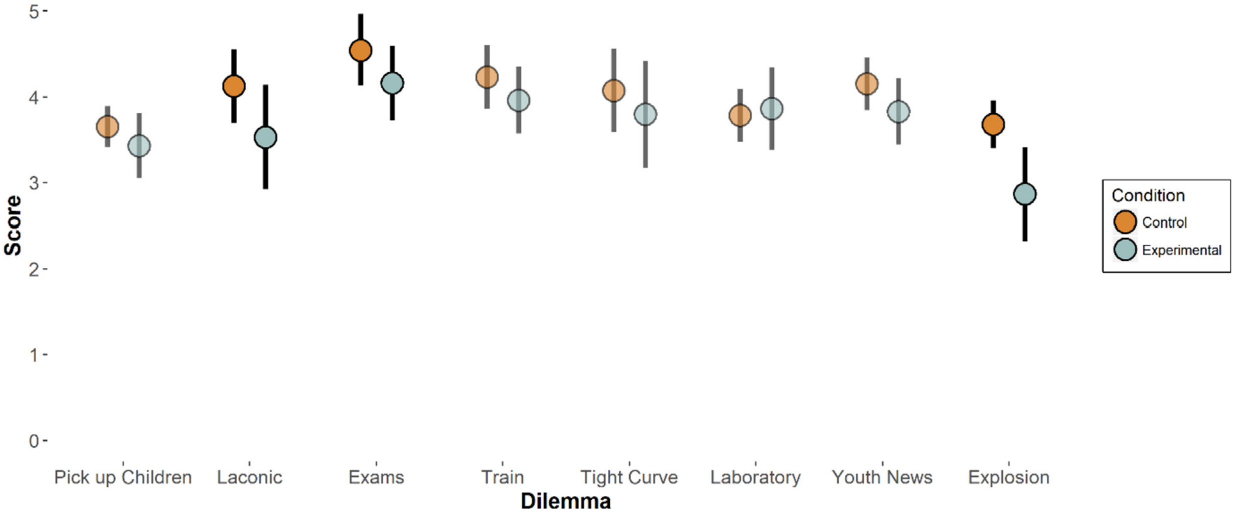


Fig. 2. Comparison of scores per dilemma between conditions.

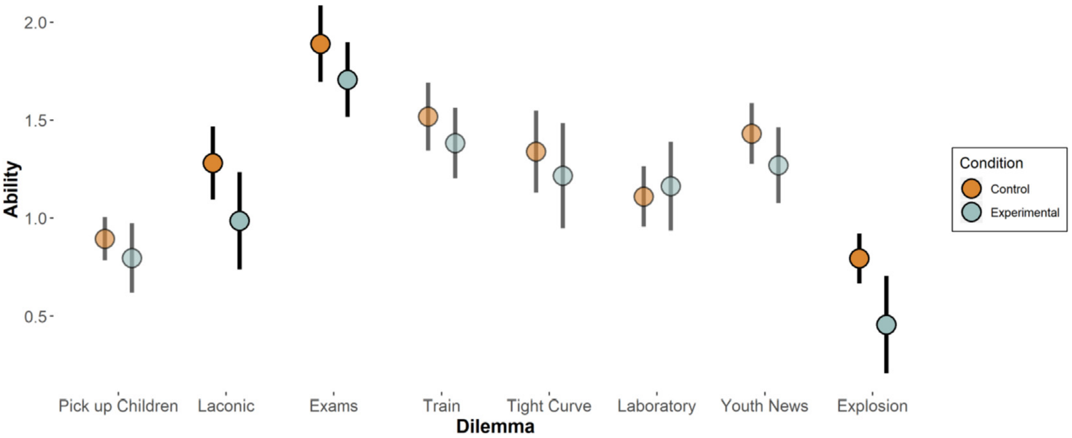


Fig. 3. Comparison of person ability parameters per dilemma between conditions.

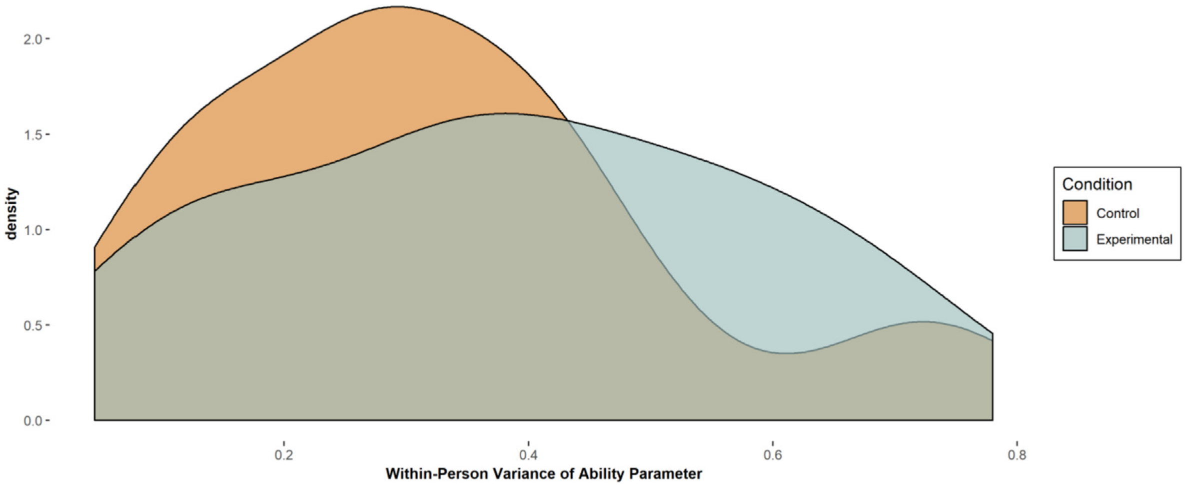


Fig. 4. Sample distributions of within-person variances in the ability parameter per condition.

relationships between within-person variances and experienced stress level.

To further explore differences in (variability of) gameplay behavior, we chose to extract more data from the game logfiles, keeping in mind that we expected differences in gameplay behavior between different levels of induced stress. One variable that could also differ between different levels of stress is the amount of time an information item was kept open and its variance. The average open time was computed per dilemma to take into account that more vague information was kept open much longer than less vague information. To correct for the difference in open time that we induced by our design (12 min available in the experimental condition for 40 information items compared to

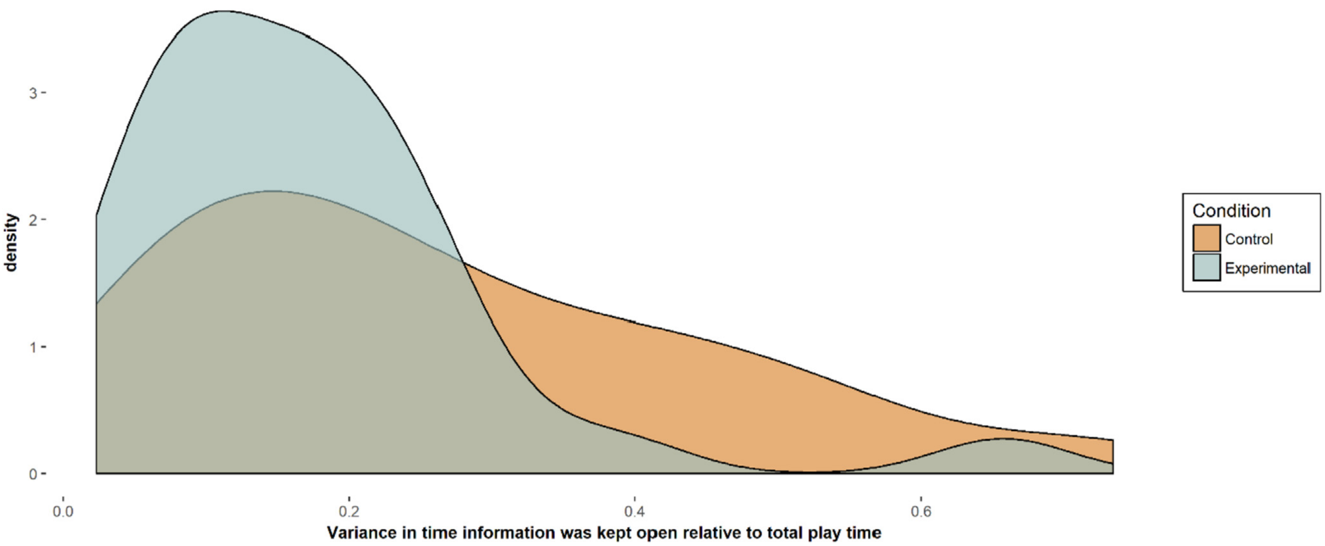


Fig. 5. Sample distributions of within-person variances of the time an information item was open per condition.

15 min in the control condition), we corrected the individual open times for the total playtime of that respective player: all open times were converted to a percentage of total time needed to finish the gameplay scenario. In this way, we took out the by-design differences between conditions and the individual working speed of players. As visualized by the greater height of the blue curve in Fig. 5, participants in the experimental condition seemed to have an overall smaller within-person variance in open times than participants in the control condition. Using an inverse-gamma regression, the difference between conditions was found to be statistically significant ($\beta_1 = 2.0696$; $t = 2.328$; $df = 66$; $p = 0.023$), indicating that participants in the experimental condition divided the available used time on each dilemma more evenly throughout the scenario.

Given that players had the opportunity to reopen already read information in order to read it again, the within-person variances of the average per dilemma counts for this were also compared between conditions. The inverse-gamma regression to test for effects of condition on the within-person variance of the average per dilemma counts of read information gave no significant result for condition effects ($\beta_1 = 4.7532$; $t = 1.533$; $df = 56$; $p = 0.131$).

3.2.2. Sensor data

Only two out of 11 measures derived from the sensor data revealed significant differences at $\alpha = 0.05$ between the conditions: the difference between the maximum and minimum skin conductance ($\beta_1 = 1.6816$; $t = 2.335$; $df = 66$; $p = 0.0226$) was bigger, and the maximum skin conductance throughout the gameplay ($\beta_1 = 1.1891$; $t = 2.241$; $df = 66$; $p = 0.0284$) was higher in the experimental condition. The minimum skin conductance showed a significant relationship with $\alpha = 0.1$ ($\beta_1 = -0.4925$; $t = 1.897$; $df = 66$; $p = 0.0623$), being lower in the experimental condition. Although only these three sensor measures differed between the two conditions, we investigated (curvi)linear relationships between sensor and in-game data. To test these relationships between sensor and in-game data, significant differences between conditions are not necessarily needed, although they would be useful in finding relevant predictors for experienced stress using in-game data.

The within-person variance computed from the person ability parameters was not a significant predictor for any of the measures identified by Holmgård et al. (2015); no linear or curvilinear relationships were found. Looking at the measures that Holmgård et al. found to be statistically significant at $\alpha = 0.1$, but not at $\alpha = 0.05$, none showed significant linear relationships with the within-person variance of the ability parameter. Neither were any curvilinear relationships found. As described, players were able to reopen already read information and read it again. Accordingly, the counts of how often information items were opened on average were correlated to the sensor features using regression analyses, as was the within-person variance of the per dilemma counts. The same was done for the average time an information item was kept open and the within-person variance of this. However, only the initial skin conductance was significantly related to the average time an information item was kept open ($\beta_1 = -2.638$; $t = -2.208$; $df = 66$; $p = 0.031$). At $\alpha = 0.1$, the last measurement of heart rate was also significantly related to the average time an information item was kept open ($\beta_1 = 8.973$; $t = 1.862$; $df = 66$; $p = 0.067$).

3.2.3. Self-report questionnaire

Out of the 80 participants, 75 participants completed the perceived arousal scale in the self-report questionnaire entirely. Four of the five participants who did not entirely complete the self-report only missed a few values. The fifth participant who did not complete the self-report was excluded from further analyses that included the self-report data. But first, the reliability of the *Perceived Arousal Scale* (PAS) was checked. Cronbach's alpha was 0.82 for that scale, and Lambda2 was 0.835, suggesting that the reliability of the Perceived Arousal Scale is satisfactory. Since the *Affect Grid* (AG) is a single-item questionnaire, reliability analyses are not applicable. Alternatively, the correlation between the AG and the PAS was tested to assess whether this scale is reliable. The AG score was not significantly correlated to the score on the perceived arousal scale ($r = 0.181$; $t = 1.578$; $df = 73$; $p = 0.119$), meaning that results based on AG scores must be viewed with caution, because since the AG did not seem to measure arousal consistently in our sample. To be complete, we still tested possible relationships involving the AG scores. Given the high reliability of the PAS, the few missing data points of the four participants who did not give a response on some items of the Perceived Arousal Scale were imputed, so that the maximum number of participants could be used in all analyses involving the self-report and gameplay data. Table 2 summarizes the mean scores and standard deviations for both scales per condition, as well as a one-tailed t-test statistic testing whether the scores are lower in the control than in the experimental condition, showing that there were no significant differences between the conditions on both scales. Note, only the self-report data of the 67 participants who completed both the scenario and the self-report was used.

The relationships between both perceived arousal scales and sensor data were also tested. A significant linear relationship was found between maximum skin conductance and the total PAS score

Table 2
Mean and standard deviations of Perceived Arousal Scale and Affect Grid scores per condition; One-tailed *t*-test statistic with fitting adjustment method for degrees of freedom.

Scale	Control:	Experimental:	*t*-test statistic
	Mean (*SD*)	Mean (*SD*)	*t* (*df*; *p*)
Perceived Arousal Scale	86.42 (10.485)	89.41 (8.317)	−1.263 (65; 0.1055)
Affect Grid	6.55 (1.350)	6.83 (0.966)	−0.930 (65; 0.1775)

($\beta_1 = -0.0536$; $t = -2.237$; $df = 64$; $p = 0.0288$), when also taking between-condition differences into account. Further, a significant linear relationship was found between skin conductance range and the total PAS total score ($\beta_1 = -0.06874$; $t = -1.999$; $df = 64$; $p = 0.0499$), also taking into account between-condition effects. No significant linear or curvilinear relationships were found between sensor data and the AG scores.

A possible relationship between the PAS and AG scores and the within-person variance in the ability parameter was also tested using regression analysis; in both cases, the results were not significant (PAS: $\beta_1 = -5.800$, $t = -1.948$, $df = 64$, $p = 0.347$; AG: $\beta_1 = -0.3853$, $t = -0.502$, $df = 64$, $p = 0.617$). Neither were any curvilinear relationships found. However, relating the within-person variance per dilemma of how often information was read to the PAS and AG scores showed higher correlations. The within-person variance of the per dilemma count of opened information was significantly correlated to the PAS score at $\alpha = 0.1$, correcting for between-condition effects ($\beta_1 = -1.691$; $t = -1.755$; $df = 64$; $p = 0.084$). Furthermore, the average count of opened information was significantly correlated to the AG score ($\beta_1 = -0.841$; $t = -2.31$; $df = 64$; $p = 0.0241$) and the within-person variance of the per dilemma count was significantly correlated to the AG score ($\beta_1 = -0.348$; $t = -3.023$; $df = 64$; $p < 0.01$), both taking into account the between-condition effects. The self-report data were neither linearly nor curvilinearly related to the average time an information item was kept open relative to the total play time, nor to the within-person variance in that measure.

4. Discussion

The general aim of this study was to find an unobtrusive measure for experienced stress in a digital serious game, by using different in-game measures. The unobtrusive measure we used for experienced stress was based on the variation in observed performance levels for analytical skills throughout the gameplay. To find indications of the validity of this way of measuring experienced stress, the gameplay data were related to condition, sensor and self-report data. While differences in analysis scores were found between the two conditions, finding an unobtrusive measure for experienced stress was less successful. The collected physiological arousal data offered only few measures that correlated with in-game measures. The chosen (simple) measures to unobtrusively assess experienced stress were, at least in our sample, not able to reliably measure the participants' stress levels.

4.1. *Effect of induced stress on in-game performance*

The first research question questioned whether we can observe a difference between the conditions with respect to the players' performance on the analyzing competency. The analyses confirmed that this is the case. Changing the game pace influenced how well the information items offered in the game were analyzed, thereby being in line with available literature suggesting that decision making can be both calm and thought-through (strategic), but also spontaneous and intuition-guided (intuitive) (Cooper, 2007; Endsley, 2000). Participants in the experimental condition indeed seemed to play the game differently than participants in the control condition. Participants in the experimental condition seemed to follow a more intuition-guided, not thought-through, decision-making strategy, which in turn might be less accurate due to a lack of situational awareness presumably caused by a heightened stress level, thereby leading to biases in judgment (Tversky & Kahneman, 1974).

Despite possibly following a more intuition-guided decision-making strategy, participants in the experimental conditions in general spent a similar amount of time on each dilemma, while participants in the control condition varied more with respect to time needed per dilemma. In the experimental condition the participants were provided with a large alarm clock and regular announcements about the amount of time left until the scenario was over; these changes could have resulted in higher awareness of time. This was particularly interesting to find, since it is in contrast to the higher within-person variances for participants in the experimental condition that we initially expected. Looking at only the participants who completed the entire scenario, the regular announcements and the large clock seemed to help to better plan ahead for the rest of the scenario, and thus proceed to the next dilemma on time, so that they were able to finish all of the upcoming dilemmas.

This more decisive, but intuitive, decision-making strategy actually is desirable up to a certain degree (Veiligheidsregio Twente, 2016). Quite often crisis managers struggle to be decisive, meaning that they tend to spend more than the necessary time on one single problem. In real-life crisis situations, the world does not stand still while the decision makers gather and analyze information. The crisis evolves, so that the crisis managers must make decisions and cannot afford to think about a single dilemma for too long. Still they must keep it in mind to base their decisions on a thorough analysis of the available information, so being too decisive also can have its downsides when information is missed or wrongly interpreted. For example, just as in real-life crisis scenarios, making decisions before all information is considered and analyzed forces the decision makers to more often follow an intuition-guided, less accurate, decision-making strategy. The available information might be vague, not directly implying the decision to make, so the decision maker could be unable to correctly interpret the main message of the information. Hence, the quality of the decision-making process may decrease when decision makers act too decisively.

Analytical skills were not measured unobtrusively, as this was not the aim of this study. It might be useful to develop an unobtrusive measure for this competency as well. In real-life crisis situations, the crisis managers do not specifically indicate what the available information implies. Crisis managers read or listen to information, interpret it, and then move on with the process. An unobtrusive measure for analytical skills would allow game developers to develop an even more realistic serious game, thereby simulating realistic situations in even greater detail. Additionally, in line with Shute's (2011) statement about keeping players in a flow state to improve the learning experience, the game could help do so by employing unobtrusive assessment methods for analytical skills.

4.2. *Unobtrusive assessment of experienced stress*

The second research question stated whether we are able to find relationships between in-game measures and (physiological) stress measures to be able to unobtrusively assess the experienced stress of a player. As described earlier, knowing the current level of experienced stress can help with improving the learning experience of a player, by adjusting the game to the player's actual state. Being able to unobtrusively assess the experienced stress level is crucial for accomplishing that in a digital serious gaming context. However, in this study the in-game measures for experienced stress were rather simple, and for the most part they did not significantly correlate with the physiological arousal data. Only the average time an information item was kept open relative to the total playtime was significantly related to the heart rate at the last measurement point and the first measurement of skin conductance. Relating the self-report data to the in-game measures also was not too insightful. Only the counts of information items read and the within-person variance of these counts were significantly correlated with participants' Affect Grid scores. Reading information less often and a smaller variance in these counts, respectively, were associated with higher Affect Grid scores. Overall, these results do not provide sufficient evidence for the self-report measures to serve as an unobtrusive measure of experienced stress, given that a significant relationship was found with only two out of 11 physiological measures.

These results could be explained through different means: First, the variance in the available sensor measures was relatively low, making it difficult to find meaningful relationships with the more variable gameplay data. The (physiological) sensor measurements differed between conditions in only three out of 11 cases. This raises a question as to whether the manipulation was effective enough to cause the desired differences between the conditions with respect to experienced stress. Second, the self-report measures were in line what the sensor data already implied. Both the Affect Grid and the Perceived Arousal Scale scores showed no significant differences between the two conditions. Third, we chose to employ rather simple methodologies. More sophisticated methods could be more effective in finding relevant indicators for the experienced level of stress. For example Bayesian networks, as a popular probabilistic method in unobtrusive assessment (e.g. Shute, 2011; Wang, Shute, & Moore, 2015), could make better use of the logged in-game data than the simple statistical models employed in this research. Additional log-data to be considered could be, for example, patterns of requesting information from only a few advisors (for example, only police services and fire services), or patterns of ignoring one advisor in particular. Fourth, based on the obtained results it seems that we also induced a direct effect on performance, just because participants in the experimental condition had less time to finish the scenario. Although this was meant to be a stress-inducing factor, we do not have a clear indication of whether this was the case, or whether the shorter time itself caused the performance drop. Checking whether this was the case could be subject of an additional experiment using two conditions with the same amount of available time. Based on our study, we can conclude that the manipulation caused differences in interacting with the game, which is a promising insight for possible future research into this topic. Last, it would be interesting to dive into similar research with active professionals in the crisis management domain, to gain insight into how the same mechanisms work in a closer-to-practice setting and also to test more sophisticated methodologies.

Declarations of interest

The authors declare that they have no conflict of interest.

Acknowledgements

This research was funded by the Netherlands Organisation for Scientific Research (NWO) program *Professional Games for Professional Skills* as part of the *Data2Game: Enhanced Efficacy of Computerised Training via Player Modelling and Individually Tailored Scenarios* research project under project number *055.16.114*. We thank Paris Mavroumostakos-Blom from Tilburg University for filtering and preparing the sensor data for our needs, and we thank Emily Fox for proofreading our manuscript.

Appendix A

Table A.1
Description of the fourth dilemma, “Train”, and the five respective information items offered.

Dilemma	Dilemma Text “Train”	The truck has caused an international train to be stranded at the station of Trouveen; because the danger zone also includes parts of the rail the train cannot go any further. There is a delegation of the “women of Srebrenica” in the train, they are on their way to the International Court in The Hague. Approximately 150 passengers are being sheltered in a sports hall close to the station, where they are provided with coffee, tea and cake.
	Question Text	Are you going to visit the sports hall to boost the morale of the stranded passengers?
Information Item	Police Services	You should remain inside the city hall. That way you are easily available for the operational leader. This is the best for the sake of this crisis.
	Fire Services	The situation is stable, but precarious. A captain has to stay on board. In other words: your presence is desired in the crisis team now.
	Public Health Services	If you would just show up in the shelter hall, that would also give a boost to the staff at the municipality.
	Legal Advisor	Nobody is indispensable. Your role in the crisis team can be fulfilled by the deputy mayor. You can afford to leave for half an hour.
	Communication and Press	It is good to show your involvement to the stranded travelers also. You have to fulfill your role as figurehead.

Appendix B

Table B.1
Correspondence markers for competency ‘Analyzing’ for response ‘Yes’.

	Communication and Press	Public health services	Fire services	Legal advisor	Police services
Pick up children	1	0	1	0	1
Laconic	0	0	1	1	1
Exams	0	0	1	0	1
Train	1	1	0	1	0
Tight curve	1	0	1	0	0
Laboratory	1	1	0	1	0
Youth news	1	0	0	1	0
Explosion	1	1	1	1	1

The table displays the correct correspondence marker, given that the response on the dilemma was ‘yes’. If the decision was ‘no’ all values are inverted. If a participant decides ‘yes’ on the dilemma ‘pick up children’, the participant should mark the information from the ‘communication and press’, ‘fire services’ and ‘police services’ advisors as corresponding, while leaving the information of the ‘public health services’ and ‘legal advisor’ unmarked. If the participant does exactly this, the resulting score would be five. Hence, if the participant does the opposite while still answering ‘yes', the resulting score would be zero. For each correctly set correspondence marker the participant receives one point. The dilemmas are scored individually. When the participant did not read an information item, the item is considered as wrongly interpreted.

Appendix C

The person ability parameters were estimated by continuously computing the marginal likelihood of the data under multiple hypotheses. For each hypothesis, the marginal likelihoods were computed using an importance sampling based procedure with 10,000 samples from the proposed parameter distribution of the posterior.

The initial proposal distribution of the ability parameter θ was a normal distribution $N(0,0.1)$. Three hypotheses were compared:

H_{a0}: $-0.1 < \theta < 0.1$

H_{a1}: $\theta > 0.1$

H_{a2}: $\theta < -0.1$

This first comparison only gives a first direction of where parameter value might be. To find the correct value, the hypotheses are adjusted and testing is repeated. Hence, if H_{a1} turned out to be the most probable hypothesis, we now would test the hypotheses:

H_{b0}: $0 < \theta < 0.2$

H_{b1}: $\theta > 0.2$

H_{b2}: $\theta < 0$

Also, the distances between the tested parameters are decreased when two hypotheses are about equally probable. This serves to find an accurate value of the parameter. Hence, later hypotheses to test could be:

H_{c0}: $0.5 < \theta < 0.55$

H_{c1}: $\theta > 0.55$

H_{c2}: $\theta < 0.5$

The procedure was repeated until the third decimal of θ was found.

References

Anderson, C. A. (1995). *Perceived arousal scale.* Retrieved from http://public.psych.iastate.edu/caa/scales/PerArous.pdf.

Anderson, C. A., Deuser, W. E., & DeNeve, K. M. (1995). Hot temperatures, hostile affect, hostile cognition, and arousal: Tests of a general model of affective aggression. *Personality and Social Psychology Bulletin, 21*(5), 434–448. https://doi.org/10.1177/0146167295215002.

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Bellotti, F., Kapralos, B., Lee, K., Moreno-Ger, P., & Berta, R. (2013). Assessment in and of serious games: An overview. *Advances in Human-Computer Interaction, 2013*. https://doi.org/10.1155/2013/136864.

Cheng, S.-Y. (2018). *Evaluation of effect on cognition response to time pressure by using EEG.* Paper presented at the conference on Advances in Human Factors and Ergonomics in Healthcare and Medical Devices. 10.1007/978-3-319-60483-1_5.

Cocea, M., & Weibelzahl, S. (2009). Log file analysis for disengagement detection in e-Learning environments. *User Modeling and User-Adapted Interaction, 19*(4), 341–385. https://doi.org/10.1007/s11257-009-9065-5.

Cohen, R. A. (2011). Yerkes–dodson law. In J. S. Kreutzer, J. DeLuca, & B. Caplan (Eds.). *Encyclopedia of clinical neuropsychology* (pp. 2737–2738). New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-79948-3 1340.

Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., & Boyle, J. M. (2012). A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education, 59*(2), 661–686.

Cooper, H. H. A. T. (2007). Decision making in a crisis. *Journal of Police Crisis Negotiations, 7*(2), 5–28. https://doi.org/10.1300/J173v07n02_02.

Crichton, M. T., & Flin, R. (2001). Training for emergency management: Tactical decision games. *Journal of Hazardous Materials, 88*(2), 255–266.

Crichton, M. T., Flin, R., & Rattray, W. A. R. (2000). Training decision makers – tactical decision games. *Journal of Contingencies and Crisis Management, 8*(4), 208–217.

Endsley, M. R. (2000). Theoretical underpinnings of situation awareness: A critical review. In M. R. Endsley, & D. J. Garland (Eds.). *Situation awareness analysis and measurement* (pp. 3–32). Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff.

Holmgård, C., Yannakakis, G. N., Martínez, H. P., Karstoft, K.-I., & Andersen, H. S. (2015). Multimodal ptsd characterization via the startlemart game. *Journal on Multimodal User Interfaces, 9*(1), 3–15. https://doi.org/10.1007/s12193-014-0160-5.

van der Hulst, A. H., Muller, T. J., Buiel, E., van Gelooven, D., & Ruijsendaal, M. (2014). Serious gaming for complex decision making: Training approaches. *International Journal of Technology Enhanced Learning, 6*(3), 249–264. https://doi.org/10.1504/ijtel.2014.068364.

Jong, W. (2017, August 13). Burgemeestersgame [Blog Post]. Retrieved from https://www.burgemeesters.nl/content/1308-burgemeestersgame.

Kim, K. H., Bang, S. W., & Kim, S. R. (2004). Emotion recognition system using short-term monitoring of physiological signals. *Medical, & Biological Engineering & Computing, 42*(3), 419–427. https://doi.org/10.1007/BF02344719.

Kowalski-Trakofler, K. M., & Vaught, C. (2003). Judgment and decision making under stress: An overview for emergency managers. *International Journal of Emergency Management, 1*, 278–289. https://doi.org/10.1504/IJEM.2003.003297.

Lopes, R., & Bidarra, R. (2011). Adaptivity challenges in games and simulations: A survey. *IEEE Transactions on Computational Intelligence and AI in Games, 3*(2), 85–99. https://doi.org/10.1109/TCIAIG.2011.2152841.

Lynch, S. M. (2007). Basics of bayesian statistics. In S. M. Lynch (Ed.). *Introduction to applied bayesian statistics and estimation for social scientists* (pp. 47–75). . Statistics for Social and Behavioral Sciences https://doi.org/10.1007/978-0-387-71265-9_3.

Mandryk, R. L., Inkpen, K. M., & Calvert, T. W. (2006). Using psychophysiological techniques to measure user experience with entertainment technologies. *Behaviour & Information Technology, 25*(2), 141–158. https://doi.org/10.1080/01449290500331156.

Maule, A., Hockey, G. J., & Bdzola, L. (2000). Effects of time-pressure on decision-making under uncertainty: Changes in affective state and information processing strategy. *Acta Psychologica, 104*(3), 283–301. https://doi.org/10.1016/S0001-6918(00)00033-0.

McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(5), 750–773. https://doi.org/10.1080/10705511.2016.1186549.

Mezey, G. (2004). Crisis management decision making. *AARMS, 3*, 267–288.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Raudys, S., & Justickis, V. (2003). Yerkes-Dodson law in agents' training. In F. M. Pires, & S. Abreu (Eds.). *Progress in artificial intelligence* (pp. 54–58). Berlin: Springer Berlin Heidelberg.

Russell, J. A., Weiss, A., & Mendelsohn, G. A. (1989). Affect Grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology, 57*(3), 493–502. Retrieved from https://doi.org/10.1037/0022-3514.57.3.493.

Sarpong, D., & Maclean, M. (2011). Scenario thinking: A practice-based approach for the identification of opportunities for innovation. *Futures, 43*(10), 1154–1163.

van de Schoot, R., Broere, J. J., Perryck, K. H., Zondervan-Zwijnenburg, M., & van Loey, N. E. (2015). Analyzing small data sets using bayesian estimation: The case of posttraumatic stress symptoms following mechanical ventilation in burn survivors. *European Journal of Psychotraumatology, 6*(1), 25216. https://doi.org/10.3402/ejpt.v6.25216.

Shi, Y., Nguyen, M., Blitz, P., French, B., Frisk, S., Torre, F., & Kumar, S. (2010). *Personalized stress detection from physiological measurements. Proceedings of the 2nd international symposium on quality of life technology*.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer Games and Instruction, 55*(2), 503–524.

Shute, V. J., & Kim, Y. J. (2014). Formative and stealth assessment. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.). *Handbook of research on educational communications and technology* (pp. 311–321). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-3185-5 25.

Sniezek, J. A., Wilkins, D. C., & Wadlington, P. L. (2001). *Jan). Advanced training for crisis decision making: Simulation, critiquing, and immersive interfaces. Proceedings of the 34th annual Hawaii international conference on system sciences*https://doi.org/10.1109/HICSS.2001.926337.

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

Susi, T., Johanneson, M., & Backlund, P. (2007). *Serious games - an overview.* Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:his:diva-1279.

T-Xchange (2018). *Mayors game.* Retrieved http://www.txchange.nl/portfolio-item/mayors-game/, Accessed date: 15 February 2018 from .

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131.

Veiligheidsregio Twente (2016). *Regionaal crisisplan Veiligheidsregio Twente - deel 1.* [Regional Crisisplan Twente Safety Region – Part 1]. Retrieved from http://www.vrtwente.nl/media/227908/regionaal-crisisplan-veiligheidsregio-twente-deel-1.pdf.

van de Ven, J. G. M., Stubbé, H., & Hrehovcsik, M. (2014). *Gaming for policy makers: It's serious!.* Cham: Games and Learning Alliancehttps://doi.org/10.1007/978-3-319-12157-4_32 Paper presented at the.

Wang, L., Shute, V., & Moore, G. R. (2015). Lessons learned and best practices of stealth assessment. *International Journal of Gaming and Computer-Mediated Simulations, 7*(4), 66–87. https://doi.org/10.4018/IJGCMS.2015100104.

Wasserman, L. (2004). *Simulation methods. All of statistics: A concise course in statistical inference.* New York, NY: Springer New York.

Yannakakis, G. N., & Hallam, J. (2008). Entertainment modeling through physiology in physical play. *International Journal of Human-Computer Studies, 66*(10), 741–755. https://doi.org/10.1016/j.ijhcs.2008.06.004.

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology, 18*(5), 459–482. https://doi.org/10.1002/cne.920180503.