Using Personalized Feedback to Enhance Cognitive Flexibility in the Context of Serious Gaming

Liselotte M. J. Eikenhout
s1475800

Master thesis (25 EC)
Cognitive Psychology and Ergonomics (CPE)

Internal supervisors:
Prof. dr. J.M.C. Schraagen
Dr. S. Borsci
External supervisor:
Dr. H.J.M. Pennings (TNO)

November 2018

UNIVERSITEIT TWENTE.

Abstract

Cognitive flexibility, as a process of adaptability, is important in the ever-changing environment. If we do not respond adaptively to changes, consequences may be severe. To address the question to what extent personalized feedback can enhance the training of cognitive flexibility in a serious game environment, we tested a PC-based decision making game and accompanying assignments. In this study, as part of a larger study, we employed a between-subjects design (conditions personalized vs. standardized) with repeated measures (four scenarios). The four scenarios were played by 46 students ($n_{pers.} = 23$, $n_{stand} = 23$), in two separate sessions with three to nine days in between. The total duration of the experiment was approximately five hours, and included several questionnaires on motivation and mental effort. During the game, the rules of the game would suddenly change unannounced. In their critical reflective thinking assignments, participants were to prioritize several actions, based on the game-play, and compare their answer to that of an expert. The expert's feedback was personalized, based on their performance, or a standardized routine answer. Several repeated-measures ANOVA's (with between-subjects factors) were performed, but no difference was found between the two conditions in adaptive performance, motivation, or mental effort. Conclusively, we must state that the personalization of feedback did not lead to a greater adaptive performance than standardized feedback in this study, probably due to the limited strength of the manipulation. Additionally, some exploratory analyses, limitations, recommendations, and implications are discussed.

Keywords: Cognitive flexibility, adaptive performance, adaptability, training, rule-change, personalized learning, personalized feedback, serious game, motivation, mental effort.

Table of Contents

Using Personalized Feedback to Enhance Cognitive Flexibility in the Context of Serious Gaming

The rapidly changing world around us requires constant adaptation, especially in learning or work environments (Bohle Carbonell, Stalmeijer, Könings, Segers, & van Merriënboer, 2014; Griffin & Hesketh, 2003; Ployhart & Bliese, 2006; Pulakos, Arad, Donovan, & Plamondon, 2000; Smith, Ford, & Kozlowski, 1997). Depending on the domain, if we do not properly respond to these changes, consequences can be severe. We may not be able to perform our job properly and be replaced as a result, or consequences can even be fatal, in a fire-fighting domain (Joung, Hesketh, & Neal, 2006) or military work environment for instance (Shadrick & Fite, 2009). It is therefore important that we look at how we can become and stay adaptive in new or changed situations. An essential and trainable component of adaptability is cognitive flexibility, which will be described in further detail below (Cañas, Fajardo, & Salmerón, 2006; Good, 2014; Mun, Oprins, Van den Bosch, Van der Hulst, & Schraagen, 2017).

As mentioned before, technological advancements have and will change the environment we work and live in, but we may use this technology to our advantage as well. New technologies offer new opportunities to train this cognitive flexibility, for example through serious gaming (Mun, Van der Hulst, et al., 2017), that is, games designed for education or training, not entertainment. Mun, Van der Hulst, et al. (2017) designed a serious game involving a complex decision-making environment, where a sudden unannounced rule-change is introduced to participants. The correct decisions made in response to this rule-change can be seen as a cognitively flexible response to this changing environment. This serious game proved effective in training cognitive flexibility (Mun, Oprins, Van den Bosch, & Schraagen, 2018); participants who were trained using rule-change scenarios adapted better to changes in the game than participants who trained using unchanging rule-scenarios.

Learning or training trajectories in general can be improved by personalization of training materials or contexts (e.g., Arroyo et al., 2014). This means that various aspects of training are adapted to the already acquired skills, preferences, and needs of the individual learner. Such adaptation of training is called personalized learning (Bulger, 2016; Van den Bosch, Peeters, & Boswinkel, 2017). Personalization of learning may therefore also be beneficial to the training of adaptability. In the present study, which is an extension of the study of Mun, Oprins, et al. (2017), participants were provided with personalized learning support to improve cognitive flexibility. The aim of the present study is therefore to enhance

the training of cognitive flexibility through personalized learning support, addressing the questions as to what extent personalized learning support enhances the training of cognitive flexibility in a serious game environment, and what the roles of motivation and mental effort are on the effectiveness of this training.

Adaptability and Cognitive Flexibility

Adaptability and cognitive flexibility are often used interchangeably. However, they are different, but interrelated concepts. *Adaptability* is a multidimensional construct that is defined as the ability to adjust effectively to new, unanticipated, and changing environments or situations (Glass, Maddox, & Love, 2013; Mun, Van der Hulst, et al., 2017; Pulakos et al., 2000; Ward et al., 2016). Where adaptability is thought to include eight dimensions (i.e., creative problem solving, dealing effectively with unpredictable and changing situations, learning new skills, knowledge, and procedures, interpersonal adaptability, cultural adaptability, dealing with emergencies, coping with stress, and physical adaptability; Pulakos et al., 2000), only some of these (i.e., creative problem-solving, dealing with emergencies, and learning new skills, knowledge, and procedures) seem to apply to cognitive flexibility directly. Similarly, different types of jobs may rely more on some dimensions of adaptability and less on others (Pulakos et al., 2000).

The definition of *cognitive flexibility* is very similar to that of adaptability, but there is a difference in specificity. Cognitive flexibility is the ability to rapidly and effectively reorganize one's knowledge structures in response to radically changed demands (Cañas et al., 2006; Glass et al., 2013; Ritter et al., 2012; Spiro, Coulson, Feltovich, & Anderson, 1988). Cañas et al. (2006) for instance, describe cognitive flexibility as a process that is dependent on attention. Cognitive flexibility requires persons to *perceive and be aware* of the changes in the environment, context, or tasks to perform. Subsequently, it requires a person to restructure their knowledge, their decision, and their plan of action accordingly. In that sense, cognitive flexibility can be considered the cognitive aspect of adaptability, while adaptability is an overarching term (Good, 2014; Mun et al., 2018). To be more precise, the earlier mentioned components of cognitive flexibility, attention and restructuring knowledge, can be compared to attention management and developing mental models, respectively, as described by Schraagen, Klein, & Hoffman (2008). They state that these processes, as supporting functions, are a means to achieve adaptability (Schraagen et al., 2008).

Another term often used in line with adaptability is *adaptive performance*. Adaptive performance describes the extent to which people perform effectively in new and complex situations (G. Chen, Thomas, & Wallace, 2005). It is suggested that adaptive performance of workers may benefit from exposure to "situations like those they will encounter on their jobs that require adaptation" (Pulakos et al., 2000, p. 623). In a similar vein, Ward et al. (2016) state that one should practice challenging problems beyond one's current abilities and should be allowed to acquire knowledge and reasoning skills from different contexts to achieve adaptive performance. So, to perform adaptively, individuals should have a high cognitive flexibility. Although its definition can differ based on the context of the research, in the present study we use adaptive performance as a measure of cognitive flexibility.

Trainability of Cognitive Flexibility. There is some inconsistency in the literature as to whether cognitive flexibility and adaptive performance are trainable (Baard, Rench, & Kozlowski, 2014). Several authors view cognitive flexibility and adaptability as a malleable skill (e.g., Cañas, Antolí, Fajardo, & Salmerón, 2005; Cañas et al., 2006; Ritter et al., 2012; Stokes, Schneider, & Lyons, 2010), whereas others argue that these constructs are innate, stable properties (e.g., Griffin & Hesketh, 2003; Ployhart & Bliese, 2006). Although not specifically mentioned, the definition of adaptive performance by Chen et al. (2005), the necessity of exposure to challenging situations by Ward et al. (2016), and the exposure to situations requiring adaptability by Pulakos et al. (2000), all suggest that adaptive performance can increase through exposure to training of that particular skill. Therefore, in the present study, cognitive flexibility is regarded as a malleable skill as well, in line with Cañas et al. (2005) and Mun et al. (2018). Mun et al. (2018) showed that participants exposed to rule-change during training sessions performed better in the test afterwards than untrained participants. This supports the results of Cañas et al. (2005), who found that when participants were trained in constant conditions they maintain strategies, while when they were trained under variable conditions they moved between strategies. So, "the type of training can affect, change or modify, to a certain degree, the cognitive flexibility or what is the same thing, the possibility that the participants adapt to the new conditions of the environment" (Cañas et al., 2005, p. 12). Also, Mun, Oprins, et al. (2017) suggested that exposure to a larger number of scenarios increased training duration, and more (adaptive) guidance may strengthen the effect of training on cognitive flexibility.

Personalized Learning

To add more adaptive guidance to the training of cognitive flexibility, one can make use of a personalized learning approach. Since cognitive flexibility depends on seeing changes in the environment and restructuring one's knowledge, training individuals in both areas should provide positive results, or adaptive responses. However, each individual learns in a different way and training focused on their specific individual needs will yield the best learning outcome and performance (i.e., higher skill level, higher learning speed, or higher learner satisfaction) for that individual (Durlach & Spain, 2014; Vaughan, Gabrys, & Dubey, 2016). Adapting learning trajectories to an individual's needs is called *personalized learning* (Bell & Reigeluth, 2014; Bulger, 2016; Goldberg et al., 2012). According to Van den Bosch et al. (2017), there are several ways to personalize learning, such as adapting the content of learning materials (e.g., assignments, feedback), adapting the presentation of the learning materials (e.g., books, articles, presentations), and the format of learning (e.g., self-study, cooperative learning). These adaptations can be made based on prior experience or performance, but also on more stable factors such as learners' characteristics or demographics.

Although theory suggests that personalized learning improves performance more so than routine, or standardized learning (e.g., Arroyo et al., 2014), empirical evidence for the effectiveness of personalized learning is still limited. Adaptations of content based on the learner's perspective can, theoretically, lead to a more suitable challenge for that learner. This is relatable to the most rapid learning within Vygotskij's zone of proximal development (Arroyo et al., 2014). Empirical evidence is rare, as Bulger (2016) for example states that "independent evaluations of the level of personalization or its efficacy in improving learning outcomes are rare" (p. 4).

Since personalized learning as a whole involves more than just adaptations on the individual level within exercises (e.g., learning format or presentation), in this study we will refer to adaptation as personalized feedback, so as to not understate the concept of personalized learning. In the study by Mun, Oprins, et al. (2017), the authors provided all participants with the same feedback in a critical reflective thinking assignment. They showed that the training of cognitive flexibility mainly relied on this assignment, while in-game performance showed little to no relation to other cognitive flexibility tasks (Mun et al., 2018). Therefore, the current study will focus on personalizing the feedback within the assignments

in realtime. That is, dynamically changing the feedback in a response to the developments during learning (Van den Bosch et al., 2017).

The use of Technology for Learning

According to Bell & Reigeluth (2014), there is a shift from structured, routine training to personalized training. Technological advancements provide new and seemingly more efficient opportunities for this type of training. For example, computer-based serious gaming (e.g., Mun, Oprins, et al., 2017), virtual worlds (e.g., Stricker & Arenas, 2013), or simulations (e.g., Cañas et al., 2005; Stokes et al., 2010) can be used to train decision making skills. So instead of having to experience real situations in which decisions can be fatal, a trainee can practice in a safe, simulated environment (e.g., at home, or at military training facilities; Shadrick & Fite, 2009).

One way to establish such a safe learning environment is *gamification*. Gamification describes the use of gaming tools for purposes of solving complex issues in various contexts, and has been applied for centuries (e.g. wargames for military strategies; E. T. Chen, 2015). An example of gamification used nowadays can be found in the flying of drones with the use of a console controller, described by E.T. Chen (2015). Gamification can have positive effects on learning if prior gaming experience and attitude towards game-based learning are taken into account (Landers & Armstrong, 2017).

Another application of gamification can be seen in the development of *serious games*. A serious game is a game designed for learning, not for entertainment, although it can still be entertaining (Ratan & Ritterfeld, 2009; Vaughan et al., 2016). Although the definition of serious games is vague (Ratan & Ritterfeld, 2009), the difference between an entertainment game and serious game primarily lies in the game designer's intentions. The increased use of serious games stems from the technological advancements allowing for more interactive instructional strategies than traditional pedagogical approaches, allowing them to be used in educational or training contexts (Ratan & Ritterfeld, 2009; Ritterfeld, Shen, Wang, Nocera, & Wong, 2009). Ritterfeld et al. (2009) provided evidence of two properties of serious games (i.e., multimodality and interactivity) contributing positively to the intended educational outcomes, that is, knowledge and know-how acquisition. Similarly, Veziridis, Karampelas, and Lekea (2017) showed that their serious game stimulated reflective thinking in ethics more so than a traditional classroom approach. Additionally, some motivational benefits were

elicited when a game environment was used, possibly increasing the likelihood of future learning in the relevant area (Ritterfeld et al., 2009).

The Present Study

In the present study, insights from the training of cognitive flexibility (e.g., Cañas et al., 2006; Mun, Oprins, et al., 2017) are combined with insights on personalized learning trajectories (e.g., Bell & Reigeluth, 2014; Bulger, 2016; Van den Bosch et al., 2017). The literature review of Van den Bosch et al. (2017) showed that personalized learning is much advocated, but rarely empirically validated. More specifically, even though cognitive flexibility is a highly valued skill, there is a gap in the exploration of personalized learning in the field of training for cognitive flexibility. Acknowledging this, we will enrich the existing empirical base with the current study. To do so, we improved the serious game designed by Mun, Van der Hulst, et al. (2017) and Mun, Oprins, et al. (2017), and added an extra scenario to increase the training duration and rule-change exposure, which is in line with their own recommendations. Continuing this research, incorporates the assumption that cognitive flexibility is a malleable skill (i.e., trainable). To personalize the learning, we provided personalized feedback between the scenarios, based on participants' adaptive performance.

This research addresses the question to what extent personalized feedback can enhance the training of cognitive flexibility in a serious game environment (RQ1). As this is an extension of the previous research by Mun, Oprins, et al. (2017), showing the effectiveness of their serious game, we will focus on the personalization of feedback and its effectiveness. We will compare learners who receive personalized feedback on their assignment, with those who receive a standard, routine answer. Based on aforementioned literature, we assume that learners in the personalized feedback group will show a steeper learning curve throughout all scenarios than learners in the standardized group in that their performance will increase at a higher rate (H1a). Additionally, we believe that those who receive personalized feedback will show a higher level of adaptive performance than those who receive standardized feedback (H1b).

Since literature on adaptability, personalized learning, and serious gaming briefly mentioned motivation as well as mental effort, the roles of these constructs will be explored and included as covariates in this research. We will address the question on the extent to which motivation and mental effort are related to each other in both conditions, and to adaptive performance (RQ2). Ritterfeld et al. (2009) showed that using gaming environments

may provide benefits for motivation, increasing the likelihood of training in the future. This is supported by Frankola (2001), who states that motivation can be critical in determining learning successes and student dropout rates in e-learning environments. Perhaps the entertaining and interactive value of games allows for them to be implemented in educational contexts as well, since they provide variation in learning. Additionally, according to Pulakos (2002), motivation is a significant predictor of adaptive performance (as cited by Ward et al., 2016). Motivation is in turn related to mental effort, as it depends on how much effort a student is willing to and has to put into the learning, and whether it will lead to a success (Paas, Tuovinen, Van Merriënboer, & Darabi, 2005). Mental effort itself is also related to cognitive flexibility, as it is required to invest mental resources into aborting automated or routine actions (Cañas et al., 2006). It seems that if the mental effort to respond in a cognitively flexible way is too high (according to the learner) or considered a waste of energy, an adaptive response will be lacking, or non-existent. Therefore, we will provide an exploration of the relation between motivation, mental effort, and adaptive performance in both conditions. We hypothesize that the personalized group will show higher levels of motivation and lower levels of mental effort than the standardized group (H2a). Also, we speculate that both motivation and mental effort are related to adaptive performance (H2b).

Method

Participants

Participants were recruited through convenience sampling and were mostly students from the University of Twente. An online participant-pool of mainly psychology students was used, as well as flyers put up throughout social sciences areas across the University of Twente, and the experimenter's social circle. A total of 46 participants completed the study (30 males), with a mean age of 21.5 (SD = 2.48, range = 18-28). Participants' nationality was mainly Dutch (52.2 %) or German (34.8 %). Participants either received course credits (n = 24) or €45,- (n = 22) as a reward for participating in both parts of the study and were randomly assigned to the personalized feedback condition (n = 23, 15 males), or standardized feedback condition (n = 23, 15 males). This study was approved by the ethics committee of the Faculty of Behavioural Sciences of the University of Twente.

Materials

Serious game. To study cognitive flexibility in a gaming environment, an adapted and improved version of the computer-based decision making game designed by Mun, Van

der Hulst, et al. (2017) was used. The current version of this decision making game consists of four scenarios (i.e., S0: Firefighter, S1: Robot war, S2: Nanotechnology, and S3: U.S. border security), increasing in difficulty. Each scenario contained a rich narrative designed for ill-structured complex decision making, and included in-game rule-changes to trigger cognitive flexibility (Figure 1).

Robot function	Assigned robot before solar storm	Assigned robot after solar storm
Hostile and armed	Red	Blue
Maintenance and unarmed	Blue	Green
Communication and unarmed	Green	Red

Figure 1. Example of rule-change from scenario 1.

The scenarios were designed to last approximately one hour each, except for S0 which had fewer rules and fewer cases and an expected duration of about forty minutes. As can be seen in Figure 2, all scenarios were similarly structured into three phases, inducing the player to: learn initial rules (i.e., learning phase), consolidate or, if not yet correctly learned, learn the initial rules (i.e., consolidation phase), and detect a sudden, unannounced rule change and learn the changed rules (i.e., test phase).

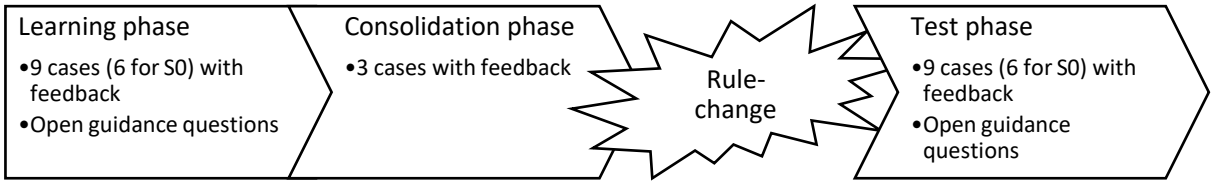

Figure 2. Graphic flow of scenario structure.

In the *learning phase* several rules were learned (i.e., two for S0 and three for S1, S2, and S3). Each rule was described by two exploratory cases and one test case. The cases consisted of a description of the situation and four options from which players chose two options each time (Figure 3). In the exploratory cases, players were exposed to the rule and all four options to choose from gave satisfactory answers. In the test case, testing whether the player understands the rule, only two out of the four options were correct. The learning phase concluded with several guidance questions, allowing players to identify relevant information. During the *consolidation phase* only one test case was presented per learned rule. If players did not comprehend the rule yet, this phase allowed for an extra opportunity to learn the initial rules. The *test phase* was identical to the learning phase in structure (i.e., three cases per rule, of which two were exploratory and one was a test case) and also included the

guidance questions. However, in contrast to the learning phase, the cases and options described the changed rules.

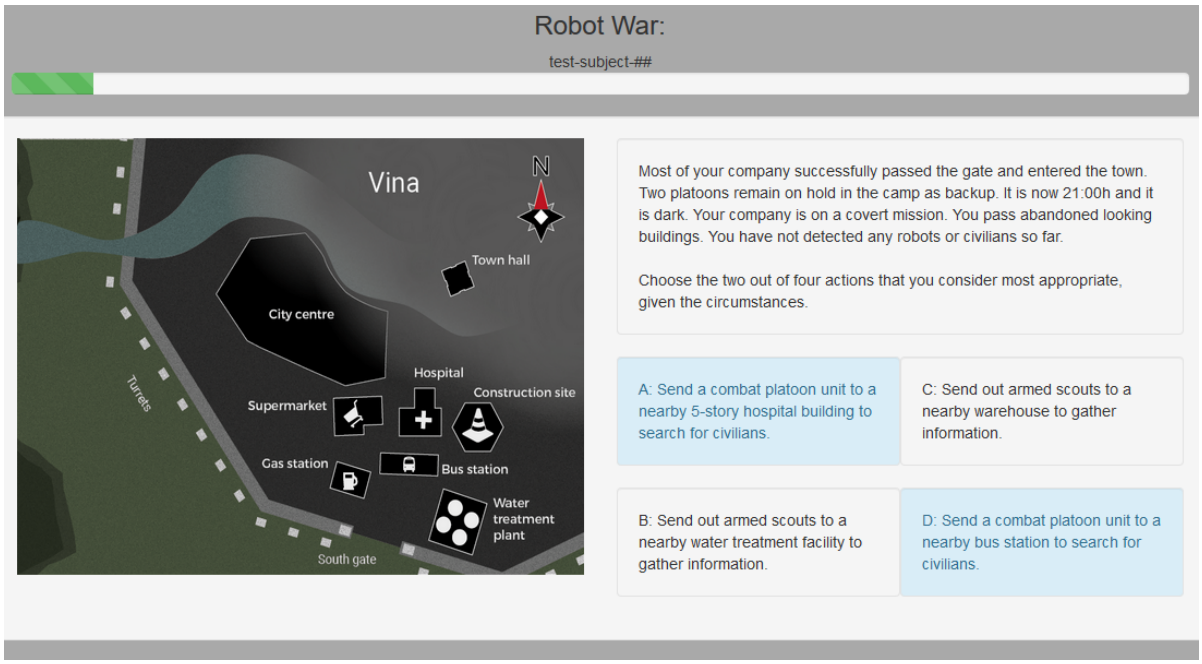

Figure 3. Screenshot of gameplay where participants select two out of four options.

Critical reflective thinking. To test whether participants had learned the initial rules and detected the rule change, a pen-and-paper based *prioritization assignment* was designed (example Appendix A). The assignment was to order four options based on suitability to a given case, and write down the reasoning behind this order. All options contained three actions, some of which were appropriate, whereas some were inappropriate. For example, one of the actions was to 'Command your combat unit to attack the green robots to prevent them from communicating with their headquarter for a backup'. Subsequently, participants received an expert's filled in assignment containing the correct answer and reasoning, which they had to compare to their own answers, writing down all the differences. The correct answer was designed as if a subject matter expert had completed the prioritization assignment, in accordance with the ShadowBox method described by Klein, Hintze, and Saab (2013). This would allow for easy comparison between the participant's and expert's answer, as well as a reference for how to do future prioritization assignments.

Types of feedback. The expert answers to the prioritization assignment after the test phase[1] differed per condition as this was the manipulation of this study (Appendix B). A total

[1] When and how the assignments were performed and feedback was distributed will be described in the procedure below.

of four types of feedback were designed, of which examples will be given below. One type was the standardized feedback, which contained no general feedback, only plain reasoning for the prioritization. The other three types (i.e., P1, P2, and P3) were designed as personalized feedback, containing both general feedback and adjusted reasoning for prioritization. P1 was most elaborate and focused on the detection of rule-change, as participants fitting this profile did not perform well in this area (Appendix B). P2 was aimed at showing the participant how to readjust their strategy after detecting the rule-changes, since participants fitting this profile struggled in this area. P3 was brief and focused mainly on motivational advice, in the sense that the participant should keep up the good work as they scored and reasoned perfectly according to the rule-changes. To limit unnecessary feedback, expert reasoning was given only for the inappropriate actions.

One of the appropriate actions in S1 was to 'Order your units to use water cannons to attack the blue robots'. An example of a standard reasoning for one of the actions is: 'Water is an effective weapon against blue robots'. In P1 this was described as: 'Just like before the solar storm, water is still an effective weapon against blue robots'. For P2 the feedback was: 'Blue robots are vulnerable to water, thus the decision to use water cannons is effective', while for P3 there was no feedback to this action as it was correct. 'The blue robots are only vulnerable to water, and cannot be destroyed with EMP grenades' is one of the reactions to an incorrect action for P3. In Appendix B, a comparison is made between the standardized feedback and the personalized feedback, P1, by marking the features excluded from the standardized feedback.

Measures

Adaptive performance. We used several measures of adaptive performance (i.e., prioritization assignments and the sum scores of the test phases). To assess the effect of the game on cognitive flexibility, we used the sum scores of the test cases in the test phases, which measured the knowledge of the changed rules. Participants could reach a maximum score of four points for each phase S0, and six points for each phase of S1, S2, and S3. In the test cases, two out of four options were correct. For each correct option, one point was awarded. For example, if options A and B were correct, and the participant chose A and C, they received one point in this case. Proportions were calculated, as the highest achievable score differed in S0.

Additionally, the scores on the prioritization assignments after the test phases were taken as measures of adaptive performance from S0, S1, and S2. In this assignment,

participants could reach a score between 8 and 16 points, depending on their prioritization. When an answer was in the correct place, four points were granted. Every place the option deviated from the correct place, one point was deducted. For example, if the correct order was A-B-C-D, but the participant switched the first two options, they were rewarded 14 points. An example of a scoring sheet for this assignment can be found in Appendix C.

Scenario 3 was differently structured as it was inherently the test scenario, in which participants did not do any expert comparison or did not receive any feedback on their performance after the learning phase. The prioritization assignment after the test phase tested the three rules separately, instead of all three rules at once, resulting in three separate scores which are not comparable to the prioritization scores of the earlier scenarios. S3's prioritization score was therefore excluded from analyses.

Motivation. The Intrinsic Motivation Inventory (IMI) is a multidimensional tool with which one can measure a participants subjective experience of an activity ("Intrinsic Motivation Inventory (IMI)," n.d.; McAuley, Duncan, & Tammen, 1989). From the IMI, two subscales were used to measure motivation four times during the experiment: The Interest/Enjoyment (7 items) and the Perceived Competence scale (6 items). The items were rated on a 7-point Likert scale, ranging from *1* (not at all true) to *7* (very true). An example item for the Interest/Enjoyment (IE) scale is "I enjoyed doing this activity very much". An example item for the Perceived Competence (PC) scale is "After working on this activity for a while, I felt pretty competent". An overview of the used items can be found in Appendix D. Each time the questionnaire was administered, the items were randomized (i.e., presented in a different order), so as to reduce bias due to order effects (Haslam & McGarty, 2003).

The reliability for both subscales of the IMI was high. The Cronbach's alpha for the IE scale across the four scenarios ranged from .90 (S0) to .94 (S2 and S3). For the PC scale, Cronbach's alpha ranged from .87 (S0) to .93 (S1). This is high, even when compared to the reliability analyses in the validation study by McAuley et al.(1989). Removing items would not yield large increases in reliability. Moreover, if items were deleted to increase reliability slightly from one of the four measurements, it would decrease reliability of another.

Mental effort. Subjective mental effort was measured using the Rating Scale of Mental Effort (RSME; Zijlstra, 1993), which was administered eight times throughout the experiment. The RSME is a 150-point vertical scale marked at 10-point intervals, including nine descriptive anchor points (i.e., absolutely no effort at 2, almost no effort at 13, a little

effort at 26, some effort at 37, rather much effort at 57, considerable effort at 72, great effort at 85, very great effort at 102, and extreme effort at 112), which are said to "refer to an underlying continuum of effort expenditure" (Zijlstra, 1993, p. 66). Participants responded by marking the scale at the point where they believed their mental effort to complete the previous task to be. Mental effort (ME) was rated a total of eight times, after each set of guidance questions, so as to account for fluctuations in different phases of the study.

Design and Procedure

In this study, an experimental, between-subjects design was employed, with repeated measures. There was one independent variable (condition), which had two levels (personalized and standardized). The dependent variable was the adaptive performance throughout each scenario. Motivation and mental effort were covariates. Participants from both conditions received the same in-game training, but received different feedback in the critical reflective thinking assignment. Both conditions were trained in rule-change during all scenarios.

Due to the length of the study, it was divided into two sessions, which were planned four to fourteen days apart. This splitting of the experiment was mainly done because of time constraints put upon the participants and their schedule. Mental fatigue could have confounded the results, had the experiment taken place in one session (Bartlett, 1941; Van der Linden, Frese, & Meijman, 2003). Also, participants now had the chance to consolidate the knowledge they had gained, and process the information. The duration of the first session was approximately 2.5 hours, and the second session lasted about 3 hours.

Figure 4 shows the timeline schematically for both day 1 and day 2. First of all, during the first session participants read a short study description (Appendix E) describing the general procedure of the experiment, after which they read and signed an informed consent conforming to the GDPR[2]. Participants were allowed to ask questions at any time (before and after signing the informed consent as well as during the experiment). If not interfering with the data collection, these questions were answered by the experimenter directly, otherwise they were answered upon completion of the experiment. After signing the informed consent, the experimenter instructed in further detail the game-play and procedure.

[2] As drawn up in cooperation with the secretary of the Ethics Committee of the Faculty of Behavioural, Management and Social Sciences at the University of Twente, Drs. L. Kamphuis-Blikman.

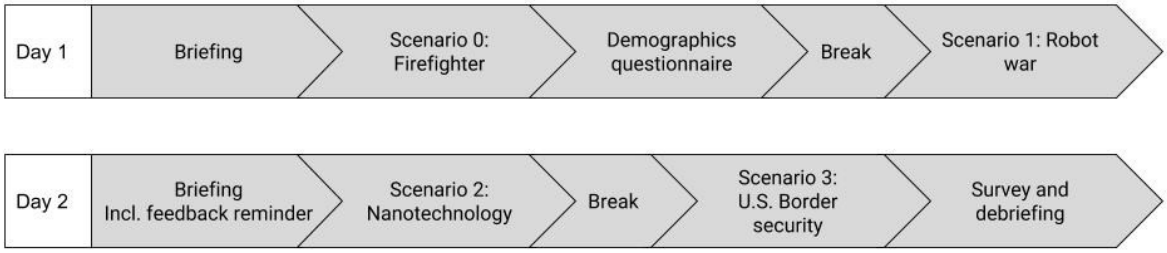

Figure 4. Timeline of the experiment.

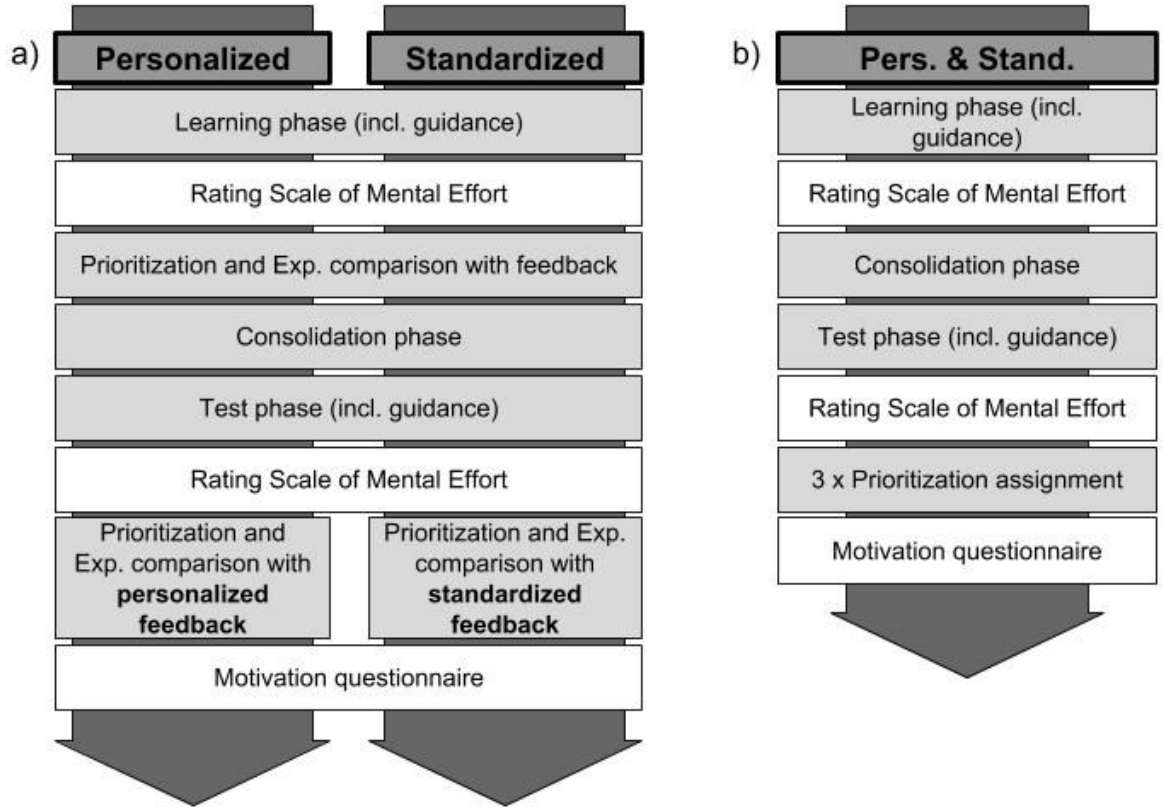

Figure 5. Detailed procedure per scenario per condition, including both training elements (light grey) and external measures (white). a) describes S0, S1, and S2. b) describes S3[3].

As can be seen in Figure 4, all participants started with the first scenario, i.e. 'S0: Firefighter', consisting of the previously described phases and assignments. For both conditions, the learning phase, first RSME, first prioritization assignment and expert comparison[4], consolidation phase, test phase, second RSME, and second prioritization assignment were identical (Figure 5a). The second expert comparison, however, differed per condition. The standardized group received a standardized expert answer to compare to their

[3] Scenario 3 was differently structured as it was inherently the test scenario, in which participants did not do any expert comparison or receive any feedback on their performance. Therefore, there was no difference between conditions.

[4] For the first prioritization assignment and comparison, both conditions received standardized feedback.

own answers and reasoning, while the personalized group received feedback and answers based on their performance on the second prioritization assignment. Two evaluators individually and simultaneously assigned participants in this condition to one of three profiles (i.e., Profile 1: does not detect rule-change; Profile 2: does not readjust strategy; or Profile 3: fully adaptive) based on their prioritization score and reasoning. When the evaluators disagreed on the profile, reasoning for profiling was discussed, and a final decision was made. Participants received expert feedback and answers based on their assigned profile (i.e., P1, P2, or P3), and were to compare the differences with their own answers and reasoning. A slight deviation between conditions was made in procedure here as well. That is, to not let the participants wait while the experimenters assigned the profile, these participants already received their break before the expert comparison. After the second expert comparison, participants filled in the motivation questionnaire and their demographics. Only the standardized group received a break after this, as the personalized group had already had their break. After the break, participants continued with the next scenario, i.e. 'S1: Robot war'. The procedure for this scenario was identical to that of S0. The first session ended after having filled in the second motivation questionnaire and checking the scheduling of the next session.

The second session followed a similar structure to the first session. The experimenter explained very briefly the procedure of the session. The standardized feedback group was allowed to directly dive into 'S2: Nanotechnology', whereas the personalized feedback group first received the general feedback from the last played scenario (i.e., S1: Robot war) as a reminder. This way, the personalized group could re-read the general tips from the previous session, and use them for the next scenario, without going into the details of the previous scenario too much. Then, also the personalized group started with S2. The procedure for S2 was identical to that of S0 and S1.

The last scenario (i.e., 'S3: U.S. Border security') was slightly different in that there was no prioritization assignment or expert comparison after the learning phase with guidance questions (see Figure 5b). After filling in the RSME, participants continued with the consolidation phase, test phase with guidance questions, and last RSME. The assignment at the end of scenario 3 consisted of three prioritization assignments, without any expert comparison. When all scenarios and assignments were done, participants filled in the last motivation questionnaire and a survey on the overall training. Finally, all participants read the debriefing handout (Appendix F), containing the actual purpose of the study and the message

that all scenarios were fictitious. They were asked to promote the study among their peers, so as to reach a wider audience, but to not discuss the content of the study.

Data analyses

Several statistical analyses were performed with IBM SPSS v23 for this study. First of all, some descriptive analyses were performed and correlations were calculated to get an overview of the data. Normality checks were performed for all measures, as well as some exploratory analyses on time between sessions and duration. For the first research question on the effect personalized or standardized feedback has on adaptive performance (RQ1; H1a and H1b), we performed two separate repeated-measures ANOVAs, in which condition (personalized or standardized) was the between-subjects factor. The within-subject factors were the repeated TP proportions (S0-S3) and prioritization scores (S0-S2). Furthermore, to determine whether both conditions showed different levels of motivation and mental effort (RQ2; H2), we performed the same repeated-measures ANOVAs with the subscale means of IE and PC, and ME ratings as within-subjects factors.

For each ANOVA performed, sphericity assumptions were tested, using Mauchly's test for sphericity. When the sphericity assumption was violated, Greenhouse-Geisser corrected tests were reported. Additionally, Levene's test was performed on all measures, to check for homogeneity of variance between both conditions.

Results

Normality checks

In Table G1, the Shapiro-Wilk outcomes are shown for all measures per scenario, and in both conditions. Tests performed on proportion test phase score and prioritization score per condition, shows that the data for adaptive performance does not portray a normal distribution. Most of the outcomes indicate a non-normal distribution. Furthermore, Q-Q plots showed large deviations from the normal distribution (for an example, see Figure G1). Still, we performed ANOVAs as they are quite robust in the face of non-normal distributions. Also, as Field (2013) states, in SPSS there is no non-parametric equivalent of the ANOVA yet. Based on the tests performed on the mean scores on interest/enjoyment, perceived competence, and mental effort per condition, we conclude that this data is approximately normally distributed. For motivation, three out of sixteen calculations showed non-normal distributions. However, none of the Q-Q plots showed large deviations from the normal distribution. For mental effort, although half of the outcomes showed non-normal

distributions, all but one of these were for the measurement during the learning phase, not the test phase. Therefore, we concluded that the data is approximately normally distributed for both groups, especially since the Q-Q plots did not show large deviations from the normal distribution. Keeping in mind that ANOVA is quite robust in cases of non-normal distributions, we assume that both motivation and mental effort are normally distributed.

Adaptive performance

Table 1 shows the descriptive statistics and correlations of all variables measured in this study (i.e., the proportion of the test phase scores, the prioritization assignment score, interest/enjoyment, perceived competence, and mental effort), divided by condition and by scenario. In Table 1 we see that participants made appropriate decisions in about two-thirds of cases in the test phases of the game (Prop; $M_{Pers} = .66$, $M_{Stand} = .62$). As for prioritization scores (Prio), we see that none of the means were below 12. This indicates an average to good performance on these assignments for both conditions, as the achievable scores ranged from 8 to 16.

The overall proportions of test phase scores and prioritization scores in each scenario per condition are visualized in Figure 6 and Figure 7, respectively. This shows that, although the average test phase score does not differ much between scenarios, the average prioritization score does. Correlations between the proportion scores in Table 1 indicate that scenarios 1, 2, and 3 were related more so than S0. When relating the proportions and the prioritization scores, a significant correlation is found in only two of the scenarios for the standardized group; S1 ($r = .54$, $p = .01$) and S2 ($r = .52$, $p < .05$). For S0 and S3 this relation in not significant, nor is it for the personalized group in any scenario.

To answer the main research question (RQ1) on whether learning curves differed between conditions, repeated measures ANOVAs were conducted on the test phase score proportions and on the prioritization assignment scores, respectively (discussed in the paragraphs below).

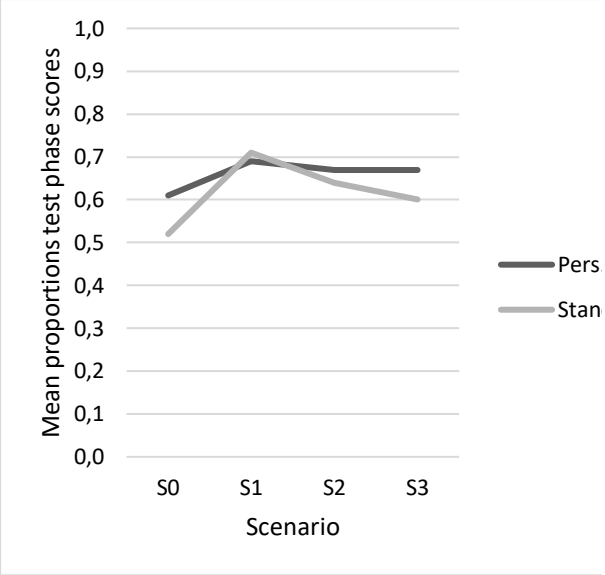

Figure 6. Average proportions of test phase scores per scenario per condition.

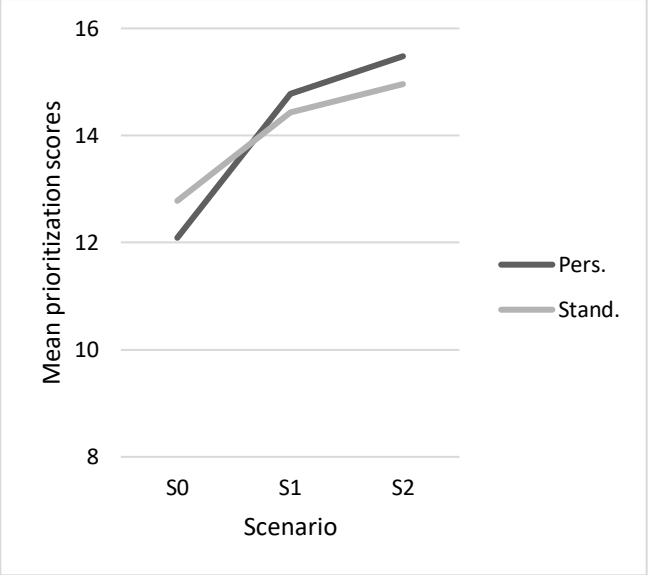

Figure 7. Average prioritization scores per scenario per condition.

Proportion test phase score. For the personalized group, the overall mean proportion of the test phases (TP) was $M = .66$ ($SD = .12$, $n = 22$, range = .42 – .88). For the standardized group this was $M = .62$ ($SD = .15$, $n = 23$, range = .19 – .83). Mauchly's test for sphericity shows that the assumption was violated for the main effects of TP proportions (χ^2 (5) = 16.66, $p < .01$), therefore Greenhouse-Geisser corrected tests are reported ($\varepsilon = .82$). The within-subjects results showed no significant interaction effect with condition (F (2.46, 105.78) = 1.36, *ns*, $\eta^2_p = .03$) nor did the between-subjects results show a significant main effect of condition (F (1, 43) = .91, *ns*, $\eta^2_p = .02$). This means that the TP scores did not differ between the personalized and standardized conditions.

Prioritization score. Overall, the mean scores on the prioritization assignment for the personalized group and the standardized group were almost identical; $M = 14.12$ ($SD = 1.54$, range = 10.67 – 16) and $M = 14.06$ ($SD = 1.73$, range = 10 – 16), respectively. Mauchly's test for sphericity shows that the assumption was violated for the main effects of prioritization score (χ^2 (2) = 9.00, $p < .05$), therefore Greenhouse-Geisser corrected tests are reported ($\varepsilon = .84$). The within-subjects results showed no significant interaction effect with condition (F (21.68, 74.03) = 1.49, *ns*, $\eta^2_p = .03$) nor did the between-subjects results show a significant main effect of condition (F (1, 44) = .01, *ns*, $\eta^2_p = .00$). This means that the prioritization score did not differ between the personalized and standardized conditions. Running the tests with S3 included did not yield different results.

Table 1

Intercorrelations, means, and standard deviations of the variables measured per scenario: Proportion TP score (Prop), Prioritization score (Prio), Interest/Enjoyment (IE), Perceived Competence (PC), and Mental Effort (ME).

		Pers. (*n*=23)	Stand. (*n*=23)	Pearson's *r*																						
		M(SD)	*M(SD)*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1.	Prop_S0	.61 (.17)	.52 (.21)	-	.19	-.40	.18	.34	**.44***	.20	.17	.37	-.14	-.13	-.25	.34	**.44***	.29	-.06	.41	**.45***	**.46***	**.44***	.11	**.44***	**.47***	**.57****
2.	Prop_S1	.69 (.16)	.71 (.19)	.26	-	**.60****	**.62****	**.42***	.34	**.50***	.29	.03	.22	-.02	-.17	**.42***	.10	.27	-.01	**.50***	.20	-.15	-.03	-.03	.00	.06	.30
3.	Prop_S2	.67 (.15)	.64 (.20)	**.42***	**.70****	-	**.47***	.07	.08	.39	.21	-.17	.26	.02	-.01	.25	-.22	.09	-.04	.14	-.07	**-.50***	**-.42***	-.25	-.40	-.38	-.14
4.	Prop_S3	.67 (.20)†	.60 (.20)	.12	**.48***	.32	-	.12	.00	.22	.17	.05	.01	.14	-.17	**.49***	-.17	.21	-.14	.27	-.10	-.40	-.09	-.32	-.21	-.17	.12
5.	Prio_S0	12.09 (2.66)	12.78 (2.75)	.28	.25	.03	.35	-	**.57****	**.52***	.28	.26	-.23	.10	.04	**.45***	.08	.09	-.03	.33	.38	.18	.20	.14	.25	.29	.41
6.	Prio_S1	14.78 (1.88)	14.43 (2.48)	.33	**.54****	.30	.03	.16	-	.21	.19	.06	.16	-.11	-.28	.15	**.45***	.07	-.24	.01	.17	-.07	-.08	.13	.10	.14	.30
7.	Prio_S2	15.48 (1.08)	14.96 (1.99)	.27	**.65****	**.52***	.32	.06	**.65****	-	**.50***	-.06	-.17	-.02	.00	.36	-.15	.32	-.08	**.47***	.38	.10	.01	-.02	-.03	.08	.08
8.	Prio_S3	13.39 (1.39)	13.10 (1.53)	.16	.00	.05	.26	.19	-.24	-.16	-	-.13	-.21	-.04	.03	.27	-.03	.29	-.03	.00	.05	-.17	-.19	-.41	.00	-.29	.21
9.	IE_S0	4.76 (.98)	4.31 (1.21)†	.21	.33	.24	.26	.25	-.03	-.28	**.47***	-	.11	.38	.34	.39	.11	.29	.37	.38	.14	.27	.31	.21	.31	**.50***	.34
10.	IE_S1	5.34 (.79)	5.25 (1.15)	.18	.13	.04	-.15	.12	**.45***	.27	.16	.00	-	.06	-.17	-.16	**.48***	-.07	-.32	-.01	.00	-.26	-.30	.24	.06	.18	.22
11.	IE_S2	4.27 (1.18)	4.14 (1.17)	.15	.01	.19	-.15	.02	-.05	-.30	.18	.35	.05	-	.30	.15	-.23	**.60****	.16	-.14	-.37	-.24	-.33	-.10	-.21	-.10	-.10
12.	IE_S3	4.33 (1.19)	3.67 (1.23)	.02	-.06	-.09	.08	.10	-.05	-.33	.34	**.48***	.00	**.55****	-	-.06	**-.51***	-.12	**.70****	-.09	-.10	.01	-.08	.17	-.03	-.06	**-.44***
13.	PC_S0	4.30 (1.20)	4.28 (.90)†	**.46***	.31	.29	**.43***	**.60****	.41	.26	.18	.33	.15	.13	.17	-	.13	.39	.25	**.49***	.10	.03	.31	-.24	-.02	.12	.36
14.	PC_S1	5.02 (.96)	5.19 (1.35)	.31	.40	.25	.10	.16	**.45****	**.66****	-.14	-.02	**.69****	-.10	-.05	**.51***	-	.10	-.24	.05	.11	.17	.18	.09	.22	.32	**.46***
15.	PC_S2	4.41 (.83)	4.52 (1.04)	.35	-.06	.26	-.13	-.28	.31	.17	-.28	-.09	.22	.40	-.09	.09	.36	-	.18	.28	-.02	.10	.06	-.23	-.11	.07	.20
16.	PC_S3	4.28 (1.05)	3.97 (.93)	.33	-.08	.16	.17	.25	-.03	-.09	.18	.24	.17	**.54****	**.48***	**.52***	.17	.38	-	.17	-.11	.21	.24	.04	-.05	.03	-.32
17.	ME1_S0	49.35 (23.88)	58.35 (23.75)	-.17	-.04	.13	-.19	-.38	-.02	.29	-.17	-.30	.11	-.33	**-.67****	-.15	.17	.13	**-.43***	-	**.73****	**.55****	**.62****	**.47***	**.55****	**.68****	**.52***
18.	ME2_S0	59.87 (23.02)	64.09 (22.20)	-.05	-.10	.22	-.12	-.34	-.08	.27	-.28	-.31	-.04	-.29	**-.58****	-.11	.13	.22	-.18	**.88****	-	**.69****	**.61****	**.54****	**.69****	**.71****	**.56****
19.	ME3_S1	66.61 (27.86)	71.87 (25.53)	.12	-.08	.03	-.26	-.16	.08	.31	-.25	**-.51***	.26	-.02	-.34	-.14	.22	.24	-.16	**.62****	**.62****	-	**.84****	**.51***	**.68****	**.67****	.34
20.	ME4_S1	67.09 (26.05)	73.65 (21.35)	.07	-.11	.08	-.20	-.05	.16	.28	**-.42***	**-.62****	.33	-.06	-.31	-.11	.31	.25	-.10	**.48***	**.56****	**.85****	-	.33	**.54****	**.63****	**.42***
21.	ME5_S2	51.70 (2.71)	65.17 (22.60)	-.18	.13	.12	.00	-.24	.01	.17	-.24	-.16	.12	-.39	**-.63****	-.13	.18	-.01	-.38	**.82****	**.76****	**.46***	.40	-	**.71****	**.72****	.17
22.	ME6_S2	62.43 (22.61)	66.74 (2.59)	-.11	.35	.30	.31	.02	.02	.26	-.05	.03	.17	-.24	-.40	.10	.23	-.12	-.18	**.60****	**.61****	.30	.34	**.85****	-	**.70****	**.64****
23.	ME7_S3	49.26 (22.86)	62.48 (27.18)	.08	.40	**.54****	.21	-.05	.02	.36	.17	.13	.13	-.09	-.31	.20	.22	-.12	-.15	**.61****	**.61****	.29	.24	**.61****	**.78****	-	**.56****
24.	ME8_S3	61.35 (28.04)	62.83 (29.82)	.04	.38	**.54****	.29	.10	.04	.35	.19	.16	.18	-.04	-.21	.23	.23	-.14	-.03	.39	**.44***	.09	.17	.41	**.71****	**.91****	-

Note. * $p < .05$ (two-tailed); ** $p < .01$ (two-tailed); † n=22, due to technical errors with the data logs, some data was missing. The top half of the Table shows the correlations for the personalized (Pers.) condition. The bottom half of the Table shows the correlations for the standardized (Stand.) condition.

Motivation and Mental Effort

In Figure 8 we see that the average interest/enjoyment rating fluctuates throughout the scenarios, with the highest rating given for S1 by both groups (M_{Pers} = 5.34, SD = .79; M_{Stand} = 5.24, SD = 1.15). A similar structure is visible for perceived competence (Figure 9). This indicates an average motivation, as both scales ranged from 1 to 7. A few questions at the end of the experiment showed that half of the participants were most excited about S1 because of the content of the scenario: "loved the scenario and it interested me the most", which is why they thought they performed best on this scenario. As for mental effort, the range of the means is fairly small (Table 1; 49.26 – 73.65), as the full range of the scale was 0-150. The results also show little variation over time, which indicates a fairly stable amount of effort asked. The label closest to the mean (62.05) is 'rather much effort' at 57.

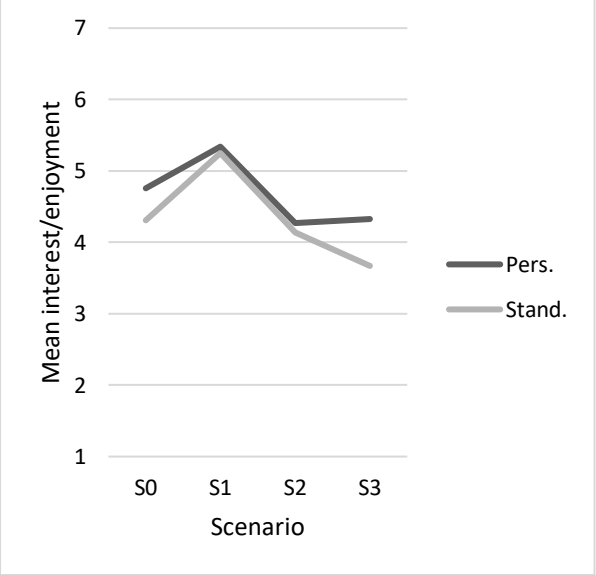

Figure 8. Interest/Enjoyment ratings per scenario per condition.

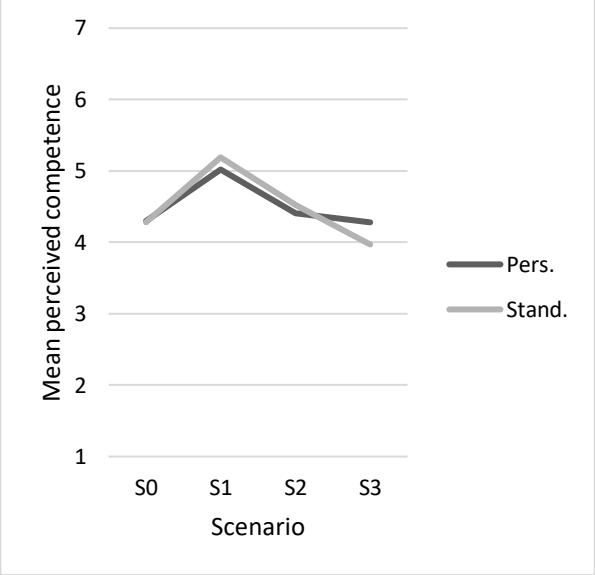

Figure 9. Perceived Competence ratings per scenario per condition.

To answer the question (RQ2) on whether motivation and mental effort were different between conditions and whether this had an effect on the training, repeated measures ANOVAs were conducted on the score on interest/enjoyment, perceived competence, and mental effort, separately. Again, time was the within-subjects factor and condition was the between-subjects factor. The correlations that were calculated, allowed us to answer the question whether adaptive performance was related to motivation and mental effort.

Interest/enjoyment. First of all, Mauchly's test showed that the assumption for sphericity was not violated for the main effects of IE (χ^2 (5) = 4.92, p = .43), therefore

sphericity was assumed. The within-subjects results showed no significant interaction effect with condition (F (3, 129) = .67, *ns*, η^2_p = .02) nor did the between-subjects results show a significant main effect of condition (F (1, 43) = 2.39, *ns*, η^2_p = .05). This means that the IE score did not differ between the personalized and standardized conditions.

Perceived competence. Again, Mauchly's test showed that the assumption for sphericity was not violated for the main effects of PC (χ^2 (5) = 6.26, p = .28), therefore sphericity was assumed. The within-subjects results showed no significant interaction effect with condition (F (3, 129) = .87, *ns*, η^2_p = .02) nor did the between-subjects results show a significant main effect of condition (F (1, 43) = .03, *ns*, η^2_p = .00). This means that the PC score did not differ between the personalized and standardized conditions.

Mental effort. In this ANOVA, all eight measurements of the RSME were the within-subject variables, while the condition was the between-subjects factor. Mauchly's test for sphericity shows that the assumption was violated for the main effects of ME (χ^2 (27) = .04, p < .001), therefore Greenhouse-Geisser corrected tests are reported (ε = .52). The within-subjects results showed no significant interaction effect with condition (F (3.62, 159.31) = .82, *ns*, η^2_p = .02) nor did the between-subjects results show a significant main effect of condition (F (1, 44) = 1.64, *ns*, η^2_p = .04). This means that, although we see that the standardized group rated their mental effort higher than the personalized group on all separate occasions (Figure 10), the ME score did not differ between the personalized and standardized conditions.

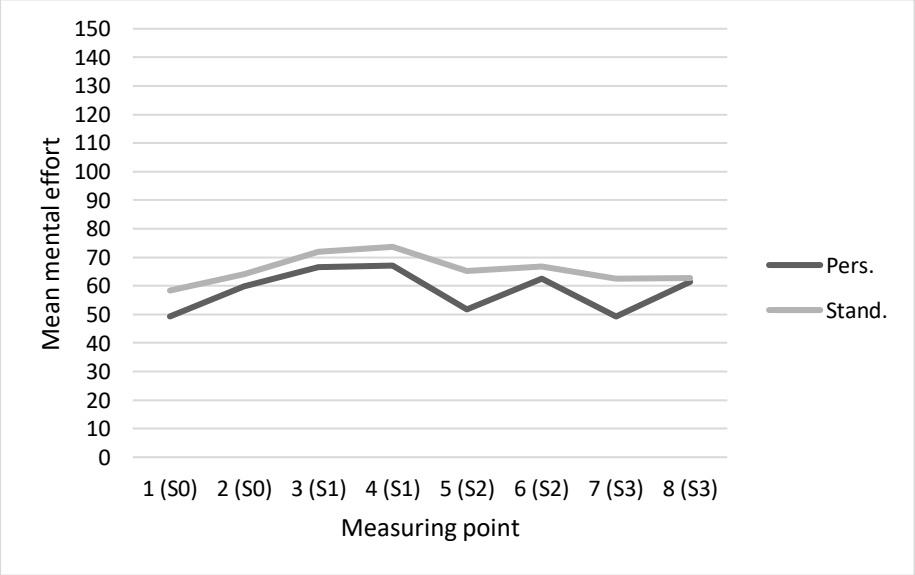

Figure 10. Mean ratings on mental effort (ME) per occasion, per condition, including the overall mean on all ratings.

Relation Adaptive Performance, Motivation, and Mental Effort

First of all, the correlations in Table 1 show that there is no association between motivation (i.e., interest/enjoyment and perceived competence) and mental effort. When time of measurement is taken into account, correlations are low to non-existent (e.g., ME2_S0 with IE_S0 and PC_S0, ME4_S1 with IE_S1 and PC_S1, etc.). The subscales of motivation, that is IE and PC, do show moderate correlations for both conditions (e.g., Pers. IE_S1 and PC_S1 $r = .48$, $p = .02$; Stand. IE_S1 and PC_S1 $r = .69$, $p < .001$), supporting the related nature of the subscales. The ME shows many significant correlations among itself, indicating that the scale and participants' rating is consistent over time.

Revisiting Table 1, we see that the proportion score was not related to IE or PC in any of the scenarios, in either group. The only significant correlation between prioritization score and IE, was found for the standardized group in S1. Therefore, a relation between IE and adaptive performance cannot be confirmed. Similarly, ME is not related to adaptive performance, as only one out of 16 correlations were significant (i.e., personalized, S0). However, PC was significantly related to the prioritization score to a certain extent, for both groups in S0 and S1 (ranging from $r = .45$ to $r = .60$), but not for S2.

Exploration of Timing and Duration

Figure 11 shows the number of days between the sessions by condition. As to be flexible towards participants, they were allowed to plan both sessions themselves. One participant completed the sessions three days apart due to unforeseen appointments, while all other participants completed both sessions between four and nine days apart. On average, the personalized group returned for the second session after 6.04 days ($SD = 1.61$), while the standardized group did so after 5.83 ($SD = 1.64$) days.

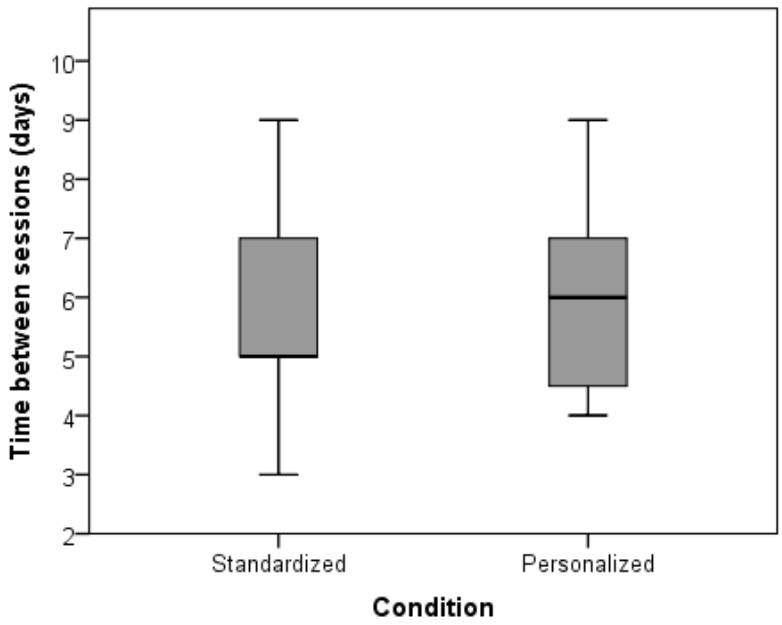

Figure 11. Boxplot of days between scenarios by condition.

Figure 12 visualizes the duration of the full experiment by condition. On average, the personalized group completed the experiment in 297.83 minutes (*SD* = 69.90), while the standardized group did so in 287.61 (*SD* = 36.24) minutes. Although the spread for the personalized group (185-435 minutes) is larger than that of the standardized group (220-355 minutes), the average duration was shorter than the approximated time for both groups, which was 5.5 hours (i.e., Pers. = 5 hours; Stand. = 4 hours and 45 minutes).

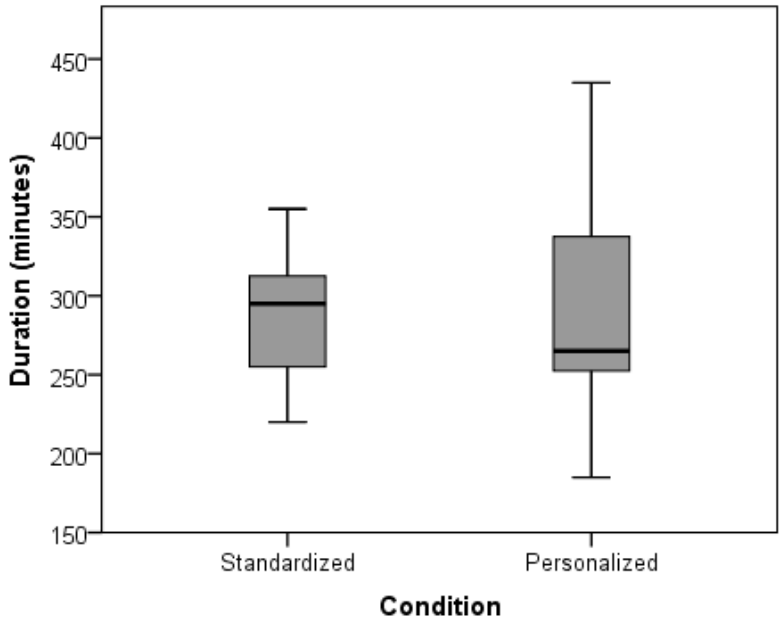

Figure 12. Boxplot of total duration of the experiment by condition.

Discussion

In this study, we examined whether personalized feedback has a more positive effect on training for cognitive flexibility than standardized feedback (RQ1). Through training sessions of approximately 5.5 hours total, we tested two groups of 23 students with the same serious game but with different types of feedback. The results described do not allow us to conclude that personalized feedback has a greater impact on training for cognitive flexibility than standardized feedback. There was no statistically significant difference between the conditions in any stage of the training. The hypotheses on the benefits of personalized feedback over standardized feedback on the learning curves (H1a) or on the adaptive performance at the end of the training (H1b), cannot be confirmed.

The additional hypotheses on the levels of motivation and mental effort in both conditions (H2a), and their relatedness to adaptive performance (H2b) were not confirmed either. Although the motivation subscales IE and PC were somewhat related, motivation was not related to mental effort, nor was it fully related to adaptive performance. Since there were no significant differences between conditions in either adaptive performance, motivation, or mental effort, not much meaningful can be said on whether personalized feedback has an effect on these constructs. For this lack of difference, there are many possible explanations. It can be related to a lack of power, a lack of variability in difficulty, or a lack of true differences in the feedback or between the participants.

The main reason for not being able to find an effect of personalized feedback, we speculate, is that the essential parts of the feedback (i.e., the correct answer) were the same in both conditions. This can lead to cherry-picking of the essential information, discarding any personalized content given. Also, if participants expected more of an experiment than a (training) game, this could lead to a more analytical view of the content and a lack of losing oneself in the game. This lack of immersion could lead to a fairly stable level of motivation and/or mental effort in the face of unexpected changes, regardless of the feedback. Perhaps giving the personalized group a differently structured feedback, or emphasizing more on reading all the content (by using check-up questions: ‘what did feedback say?’) will allow for greater immersion and differences in performance, motivation, and/or mental effort. If participants truly noticed the personalized or standardized feedback and still no differences are found, then hypotheses on better learning through personalized feedback should be revisited.

As for the exploratory analyses performed on time between sessions and duration, not much difference was found between the conditions. One can assume that the time between sessions is relevant for performance, in that participants forget relevant information or cannot consolidate their knowledge. However, on time between scenarios, both groups showed a similar spread and mean at around six days. This indicates that both groups had an equal chance of noise due to forgetting or lack of consolidation. For duration, the spread of the personalized group was larger than of the standardized group, while the mean duration was similar again. This is probably due to the differences in reading and learning speed. As performance did not differ much between conditions in this study, these measures cannot help us explain much. However, in future research these measures may help in explaining differences between the two conditions.

General Remarks

Some other interesting remarks can be made about the results and this study in general. For instance, there was a visible trend in performance considering the prioritization score. In S0, participants scored lowest, then performance increased for both S1 and S2. This indicates some type of learning effect during these scenarios. Although speculative, the performance in the first phase of the experiment (i.e., S0) could have been affected by a novelty effect. It may also indicate the content or richness of S0 deviates too much from the other scenarios, as the number of rules learned differed.

Considering that a learning curve would suggest reaching a plateau, or resulting in a ceiling effect, the results from S3 are odd. We believe the drop in performance in S3 is mostly due to the lack of comparability with the other scenarios. The lack of feedback and the lack of critical reflective thinking assignment after the learning phase, may have decreased the learning opportunity for participants. Combined with the increased difficulty and complexity of the last assignments, which asked for a different level of understanding than the other scenarios, this may have reduced the learning effects. Synchronizing S3's structure with the other scenarios could tell us whether a ceiling effect is present for adaptive performance within this experiment. That is, we can tell whether this part of cognitive flexibility can be trained further, or whether there is a limit to how cognitively flexible someone can become.

Also, the measures of adaptive performance, that is proportion test phase score and prioritization score, were related in some stages of the training, but not all. This indicates that just the game score, or just the critical reflective thinking assignment might not be enough as

a measure of cognitive flexibility. The game and assignments focus on the rule-change adaptations as an aspect of cognitive flexibility. In a previous study by Mun et al. (2018), the score on the critical reflective thinking assignment (i.e., prioritization score) was most representative of cognitive flexibility when compared to external measures. However, the adaptations made to the scenarios, in comparison to all previous versions (Mun, Van der Hulst, et al., 2017; Mun et al., 2018; Mun, Oprins, et al., 2017), may lead to a lack of comparability of the results.

Limitations and Recommendations

One of the most important limitations is that the manipulation was not strong enough, in that the personalized feedback and the standardized feedback were too similar. As mentioned before, the correct answer was the same for both groups and cherry-picking of the essential information may have led to the lack of difference between groups. For instance, it was only necessary to notice the terms 'blue robot', 'not vulnerable' or 'ineffective', and 'EMP' to see that this action was ineffective. The whole story behind why the rules changed or did not change is then irrelevant to the participant. This indicates that the personalization was, in effect, not different from the standardized feedback enough. Therefore, we propose to have another look at the personalization and presentation of feedback to really differentiate between the two conditions, perhaps with experts in this area.

Another limitation of this study is the homogeneity of the sample that was taken. Even though the nationalities and study majors were quite diverse, all participants were fairly highly educated. This may have produced a selection bias, hence results are not generalizable to other populations than were sampled. Our recommendation for future research is therefore to not only test University students, but to test a more heterogeneous sample, including people from different domains and backgrounds.

Also, although the prioritization assignment is thought to predict adaptive performance in that it tests perception and re-strategizing after sudden rule-changes, it does not grasp the full extent of cognitive flexibility or adaptability. Therefore, in future studies, we advise to incorporate other aspects of cognitive flexibility than merely rule-change.

Due to time constraints and lack of manpower, we could not test all the conditions we had hoped to test. Originally, the plan was to include a third group (i.e., control group) who would receive no critical reflective thinking assignment, nor any feedback at all. We would have been able to compare both the experimental groups (personalized and standardized) with

this control group. This would have allowed the possibility to make statements about critical reflective thinking in general, and whether this type of assignment is beneficial for the training of cognitive flexibility. Future research should therefore aim to test a control group under the same conditions as this study's experimental groups.

Implications

A few remarks can be made looking at the previously discussed literature and the results of this study. First of all, although the current study does not prove cognitive flexibility to be trainable to a certain extent, we assumed it to be as this is an extension of (Mun et al., 2018). We did see a learning curve developing in the scenarios (i.e., S0 – S2), but we did not find a plateau or ceiling effect. This implies that either the duration or the difficulty of the training was not extensive enough to show us the true learning curve of cognitive flexibility in this training.

Additionally, as mentioned before, many adaptations were made to the scenarios' content. The statement that adaptive performance, or prioritization score, in this study is representative of some measures of cognitive flexibility can therefore be questioned. The format of the prioritization assignments did not change, which is why this argument can be contested. However, a pre-test post-test construction comparing aspects of cognitive flexibility to the adaptive performance measured can clarify the representation of these adaptive performance measures as such.

Furthermore, in accordance to Vygotskij's zone of proximal development (as mentioned by Arroyo et al., 2014), the feedback was personalized based on the previous performance of participants. The fact that we did not find any significant differences between the conditions, can be due to a lack of differentiation between the personalized and standardized feedback, which may have been affected by the aforementioned cherry-picking as well the design of the feedback. Also, Van den Bosch et al. (2017) mention several ways in which learning can be personalized, but we only personalized in one way (i.e., learning content). Perhaps a more elaborate, more diverse way of personalization (e.g., delivering the feedback orally instead of through text) can induce a greater learning effect.

Another way to change the delivery of feedback, is to digitalize more of the assignments. As mentioned by Ritterfeld et al. (2009), a serious game should be interactive to induce motivation and immersion. The switching between screen and pen-and-paper assignments could have obstructed immersion into to the game. Therefore, digitalizing the

prioritization assignments and feedback, and integrating them with the serious game may increase the motivation and performance.

Conclusion

In the end, we may state that the personalization of feedback did not lead to a greater adaptive performance than standardized feedback in this study. However, to help ourselves to adapt to unexpected changes and cope with the ever-changing world around us, cognitive flexibility as a supporting process of adaptability, and its trainability should be further examined.

Acknowledgements

I extend my gratitude to Heleen Pennings, Jan Maarten Schraagen, Simone Borsci, Yelim Mun, Esther Oprins, Karel van den Bosch, Hester Stubbé-Alberts, and Sander Koning for their feedback, support, and/or supervision during my graduation process. Additionally, I would like to thank the TPI department of TNO for allowing me an internship through which I have grown as a person and a professional.

References

Arroyo, I., Woolf, B. P., Burelson, W., Muldner, K., Rai, D., & Tai, M. (2014). A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect. *International Journal of Artificial Intelligence in Education*, *24*(4), 387–426. https://doi.org/10.1007/s40593-014-0023-y

Baard, S. K., Rench, T. A., & Kozlowski, S. W. J. (2014). Performance adaptation: A theoretical integration and review. *Journal of Management*, *40*(1), 48–99. https://doi.org/10.1177/0149206313488210

Bartlett, F. C. (1941). Fatigue following highly skilled work. *Nature*, *147*, 717–718. https://doi.org/10.1038/147717a0

Bell, H. H., & Reigeluth, C. M. (2014). Paradigm change in military education and training. *Educational Technology*, 52–57.

Bohle Carbonell, K., Stalmeijer, R. E., Könings, K. D., Segers, M., & van Merriënboer, J. J. G. (2014). How experts deal with novel situations: A review of adaptive expertise. *Educational Research Review*, *12*, 14–29. https://doi.org/10.1016/j.edurev.2014.03.001

Bulger, M. (2016). *Personalized learning: The conversations we're not having*. Retrieved from https://datasociety.net/pubs/ecl/PersonalizedLearning_primer_2016.pdf

Cañas, J. J., Antolí, A., Fajardo, I., & Salmerón, L. (2005). Cognitive inflexibility and the development and use of strategies for solving complex dynamic problems: effects of different types of training. *Theoretical Issues in Ergonomics Science*, *6*(1), 95–108. https://doi.org/10.1080/14639220512331311599

Cañas, J. J., Fajardo, I., & Salmerón, L. (2006). Cognitive flexibility. In *International encyclopedia of ergonomics and human factors* (Vol. 1, pp. 297–301). CRC Press Boca Raton, FL.

Chen, E. T. (2015). The Gamification as a resourceful tool to improve work performance. In T. Reiners & L. C. Wood (Eds.), *Gamification in education and business* (pp. 473–488). Springer. https://doi.org/10.1007/978-3-319-10208-5_24

Chen, G., Thomas, B., & Wallace, J. C. (2005). A multilevel examination of the relationships among training outcomes, mediating regulatory processes, and adaptive performance. *Journal of Applied Psychology*, *90*(5), 827–841. https://doi.org/10.1037/0021-

9010.90.5.827

Durlach, P. J., & Spain, R. D. (2014). *Framework for Instructional Technology: Methods of Implementing Adaptive Training and Education*. U.S. Army Research Institute for the Behavioral and Social Sciences. Retrieved from http://www.dtic.mil/docs/citations/ADA597411

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. SAGE Publications Sage UK: London, England.

Frankola, K. (2001). Why online learners drop out. Retrieved July 29, 2018, from http://www.workforce.com/2001/06/03/why-online-learners-drop-out/

Glass, B. D., Maddox, W. T., & Love, B. C. (2013). Real-time strategy game training: emergence of a cognitive flexibility trait. *PLoS One*, *8*(8), 1–7. https://doi.org/10.1371/journal.pone.0070350

Goldberg, B., Brawner, K., Sottilare, R., Tarr, R., Billings, D. R., & Malone, N. (2012). Use of evidence-based strategies to enhance the extensibility of adaptive tutoring technologies. In *Proceedings of The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC) 2012* (pp. 1–12). Orlando, FL. Retrieved from https://gifttutoring.org/attachments/download/145/12288.pdf

Good, D. (2014). Predicting real-time adaptive performance in a dynamic decision-making context. *Journal of Management and Organization*, *20*(6), 715–732. https://doi.org/10.1017/jmo.2014.54

Griffin, B., & Hesketh, B. (2003). Adaptable behaviors for successful work and career adjustment. *Australian Journal of Psychology*, *55*(2), 65–73. https://doi.org/10.1080/00049530412331312914

Haslam, S. A., & McGarty, C. (2003). *Research methods and statistics in psychology*. London, England: Sage.

Intrinsic Motivation Inventory (IMI). (n.d.). Retrieved from http://selfdeterminationtheory.org/intrinsic-motivation-inventory/

Klein, G., Hintze, N., & Saab, D. (2013). Thinking inside the box: The ShadowBox method for cognitive skill development. In *Proceedings of the 11th International Conference on Naturalistic Decision Making* (pp. 121–124). Marseille, France.

Landers, R. N., & Armstrong, M. B. (2017). Enhancing instructional outcomes with gamification: An empirical test of the Technology-Enhanced Training Effectiveness Model. *Computers in Human Behavior*, *71*, 499–507. https://doi.org/https://doi.org/10.1016/j.chb.2015.07.031

McAuley, E., Duncan, T., & Tammen, V. V. (1989). Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research Quarterly for Exercise and Sport*, *60*(1), 48–58. https://doi.org/10.1080/02701367.1989.10607413

Mun, Y., Oprins, E., Van den Bosch, K., & Schraagen, J. M. (2018). *Game-based training to stimulate learners' adaptation process: Effects on adaptive performance*. Manuscript in preparation.

Mun, Y., Oprins, E., Van den Bosch, K., Van der Hulst, A., & Schraagen, J. M. (2017). Serious gaming for adaptive decision making of military personnel. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 61, pp. 1168–1172). https://doi.org/10.1177/1541931213601776

Mun, Y., Van der Hulst, A., Oprins, E., Jetten, A., Van den Bosch, K., & Schraagen, J. M. (2017). Serious Gaming Design for Adaptability Training for Military Personnel. *Journal of Cyber Security and Information Systems*, *5*(4), 16–23.

Paas, F., Tuovinen, J. E., Van Merriënboer, J. J. G., & Darabi, A. A. (2005). A motivational perspective on the relation between mental effort and performance: Optimizing learner involvement in instruction. *Educational Technology Research and Development*, *53*(3), 25–34. https://doi.org/10.1007/BF02504795

Ployhart, R. E., & Bliese, P. D. (2006). Individual adaptability (I-ADAPT) theory: Conceptualizing the antecedents, consequences, and measurement of individual differences in adaptability. In *Understanding adaptability: A prerequisite for effective performance within complex environments* (pp. 3–39). Emerald Group Publishing Limited.

Pulakos, E. D., Arad, S., Donovan, M. A., & Plamondon, K. E. (2000). Adaptability in the workplace: Development of a taxonomy of adaptive performance. *Journal of Applied Psychology*, *85*(4), 612–624. https://doi.org/10.1037//0021-9010.85.4.612

Ratan, R., & Ritterfeld, U. (2009). Classifying Serious Games. In U. Ritterfeld, M. Cody, & P. Vorderer (Eds.), *Serious Games: Mechanisms and Effects* (pp. 10–24). New York: Routledge.

Ritter, S. M., Damian, R. I., Simonton, D. K., van Baaren, R. B., Strick, M., Derks, J., & Dijksterhuis, A. (2012). Diversifying experiences enhance cognitive flexibility. *Journal of Experimental Social Psychology*, *48*(4), 961–964. https://doi.org/10.1016/j.jesp.2012.02.009

Ritterfeld, U., Shen, C., Wang, H., Nocera, L., & Wong, W. L. (2009). Multimodality and interactivity: Connecting properties of serious games with educational outcomes. *CyberPsychology & Behavior*, *12*(6), 691–697. https://doi.org/10.1089/cpb.2009.0099

Schraagen, J. M., Klein, G., & Hoffman, R. R. (2008). The macrocognition framework of naturalistic decision making. In J. M. Schraagen, L. G. Militello, T. Ormerod, & R. Lipshitz (Eds.), *Naturalistic Decision Making and Macrocognition* (pp. 3–25). Ashgate Publishing Limited Aldershot.

Shadrick, S. B., & Fite, J. E. (2009). *Assessment of the Captains in Command training program for adaptive thinking skills*. U.S. Army Research Institute for the Behavioral and Social Sciences. Retrieved from http://www.dtic.mil/docs/citations/ADA507445

Smith, E. M., Ford, J. K., & Kozlowski, S. W. J. (1997). Building adaptive expertise: Implications for training design strategies. *Training for a Rapidly Changing Workplace: Applications of Psychological Research*, 89–118. https://doi.org/10.1037/10260-004

Spiro, R. J., Coulson, R. L., Feltovich, P. J., & Anderson, D. K. (1988). *Cognitive Flexibility Theory: Advanced Knowledge Acquisition in Ill-Structured Domains*. Champaign, IL: Center for the Study of Reading.

Stokes, C. K., Schneider, T. R., & Lyons, J. B. (2010). Adaptive performance: A criterion problem. *Team Performance Management: An International Journal*, *16*(3/4), 212–230. https://doi.org/10.1108/13527591011053278

Stricker, A., & Arenas, F. (2013). Gamification Strategies for Developing Air Force Officers. Retrieved July 29, 2018, from https://www.learningsolutionsmag.com/articles/1190/gamification-strategies-for-developing-air-force-officers

Van den Bosch, K., Peeters, M. M. M., & Boswinkel, R. A. (2017). *Literature review on individual learning concepts. Part A: Personalized Learning*. Soesterberg, The Netherlands.

Van der Linden, D., Frese, M., & Meijman, T. F. (2003). Mental fatigue and the control of cognitive processes: effects on perseveration and planning. *Acta Psychologica*, *113*(1), 45–65. https://doi.org/10.1016/S0001-6918(02)00150-6

Vaughan, N., Gabrys, B., & Dubey, V. N. (2016). An overview of self-adaptive technologies within virtual reality training. *Computer Science Review*, *22*, 65–87. https://doi.org/10.1016/j.cosrev.2016.09.001

Veziridis, S., Karampelas, P., & Lekea, I. (2017). Learn by playing: A serious war game simulation for teaching military ethics. In *IEEE Global Engineering Education Conference (EDUCON), 2017* (pp. 920–925). https://doi.org/10.1109/EDUCON.2017.7942958

Ward, P., Hutton, R., Hoffman, R., Gore, J., Anderson, T., & Leggatt, A. (2016). *Developing skilled adaptive performance: A scoping study*. Yeovil, UK: BAE Systems.

Zijlstra, F. R. H. (1993). *Efficiency in work behaviour: A design approach for modern tools*. Technical University Delft, The Netherlands. Retrieved from https://repository.tudelft.nl/islandora/object/uuid:d97a028b-c3dc-4930-b2ab-a7877993a17f/?collection=research

Appendix A

Prioritization assignment scenario 1, test phase.

Decision making 1.2

Although you won the battle of Vina, not all robots in Vina were destroyed. They retreated and you have not received any intel about enemy robots until this morning. This morning, you received a report by the scout, about the location of the robots. Your scouts discovered a factory where new robots are being built. This factory is located about 20 km South of Vina. Your scouts discovered that the factory is secured by armed robots and there are turrets installed at every entrance of the factory. Based on this information you deployed your units near the factory, outside the turrets' radars. It is your mission to destroy the robots and the factory to terminate the construction of new robots. You are preparing an attack.

Below are a series of possible actions you may take. Some options describe a set of highly suitable actions, other options show less suitable actions, or even not suitable at all.
Given from what you have learned during this scenario, order the options in terms of their suitability. For every option, write down a reason why you put the particular option in 1st place, 2nd place etc. on the suitability ranking (min. 1 sentence, max. 3 sentences for every option).

Most suitable option	2nd best option	3rd best option	Least suitable option

A. - Command your units to send drones with cameras into the factory and order your units to pass the turrets on foot (via the passage where the turret's radar cannot detect them).
- Attack the blue robots when entering the factory, so they cannot fire back.
- Order your units to use EMP grenades to attack the blue robots.

B. - Command your units to send tanks to pass the turrets and to enter the factory.
- Attack the green robots when entering the factory, so they cannot fire back.
- Order your units to use water cannons to attack the green robots.

C. - Command your units to pass the turrets on foot (via the passage where the radar of turrets cannot detect them).
- Attack the red robots when entering the factory, so they cannot fire back.
- Order your units to use water cannons to attack the red robots.

D. - Command your units to send drones with cameras into the factory and to pass the turrets on foot (via the passage where the turrets' radars cannot detect them).
- Attack the blue robots when entering the factory, so they cannot fire back.
- Order your units to use water cannons to attack the blue robots.

Appendix B

Expert answer for comparison of the prioritization assignment for scenario 1, test phase for P1. The highlighted parts were not included in the standardized edition. Also, phrasing was different between the conditions, but key aspects of the reasoning (i.e., the correct answer) were the same for both conditions.

Expert answer 1.2 – P1

Below, along with some general feedback, the answers of the subject matter expert are provided. It consists of the ranking and reasoning, just like you have made yourself. You have to compare the order and reasoning of the expert with your own order and reasoning. Write down the differences between your own ordering and reasons and those of the subject matter expert in the table you were given.

General feedback

After the solar storm, the situation was different. The solar storm affected the behaviors and functions of the turrets and robots. For instance, blue robots are now armed, while they were meant for maintenance before the solar storm. If you take into account these differences, you could improve your decision making, so as to match it more with the expert's answers and reasoning. After that, you could carefully adjust your strategies according to the different behaviors and functions of the turrets and robots. This will help you to better deal with the situation after the solar storm.

Expert answer and reasoning

Most suitable option	2nd best option	3rd best option	Least suitable option
D	A	C	B

D **This is indeed the best option, because:**

- Before the solar storm, turrets could be avoided on foot only, via the passage where the radars of turrets have no coverage. After the solar storm, turrets can be avoided by flying over (e.g., the bird safely flying over the turret) as well as on foot.
- After the solar storm, blue robots are armed (i.e., now they carry weapons). Therefore, they should be attacked first, as they can attack your units. Before the solar storm, the blue robots were for maintenance.
- Just like before the solar storm, water is still an effective weapon against blue robots.

A **This is the second best option, because:**

- After the solar storm, turrets can be avoided by flying over (e.g., the bird safely flying over the turret) as well as on foot, via the passage where the radars of turrets have no coverage. In comparison, before the solar storm, turrets could be avoided on foot only.
- Before the solar storm, the blue robots were for maintenance. Now, after the solar storm, blue robots are armed (i.e., now they carry weapons). Therefore, they should be attacked first, as they can attack your units.

- The blue robots are *still only* vulnerable to water, and *cannot* be destroyed with EMP grenades.

C **This is the third best option, because:**

- Before the solar storm, turrets could be avoided on foot only, via the passage where the radars of turrets have no coverage. After the solar storm, turrets can be avoided by flying over (e.g., the bird safely flying over the turret) as well as on foot.
- After the solar storm, the red robots should *not* be attacked first, because the red robots are for communication (e.g., carrying communication devices), while the blue robots are armed. In comparison, before the solar storm, the red robots were armed while the blue robots were for maintenance.
- The red robots *still cannot* be destroyed with water, *only* with EMP grenades.

B **This is the least preferred option, because:**

- The turrets *still cannot* be approached using vehicles (e.g., tanks), because after the solar storm, turrets can only be avoided safely by flying over and on foot.
- After the solar storm, green robots are for maintenance and not for combat (e.g., carrying tools). Therefore, they do not pose a threat to your units. Before the solar storm, the green robots were for communication.
- The green robots *cannot* be destroyed using water, *still* the only way to attack green robots is hacking them.

Appendix C

Scoring sheet for prioritization assignment scenario 1, test phase.

Decision point 1.2 answer: DACB

	Possible answer	Scoring (min 8-max 16)
1	DACB	16
2	DABC	14
3	DCAB	14
4	ADCB	14
5	DBAC	12
6	DBCA	12
7	DCBA	12
8	CADB	12
9	CDAB	12
10	ACDB	12
11	ADBC	12
12	ABCD	10
13	ABDC	10
14	ACBD	10
15	BACD	10
16	BADC	10
17	BDAC	10
18	BDCA	10
19	CABD	10
20	CDBA	10
21	BCAD	8
22	BCDA	8
23	CBAD	8
24	CBDA	8

Appendix D

Used subscales from the Intrinsic Motivation Questionnaire.

For each of the following statements, please indicate how true it is for you, using the following scale:

1	2	3	4	5	6	7
(not at all true)			(somewhat true)			(very true)

Interest/Enjoyment:

1. I enjoyed doing this activity very much
2. This activity was fun to do.
3. I thought this was a boring activity. (R)
4. This activity did not hold my attention at all. (R)
5. I would describe this activity as very interesting.
6. I thought this activity was quite enjoyable.
7. While I was doing this activity, I was thinking about how much I enjoyed it.

Perceived Competence:

1. I think I am pretty good at this activity.
2. I think I did pretty well at this activity, compared to other students.
3. After working at this activity for awhile, I felt pretty competent.
4. I am satisfied with my performance at this task.
5. I was pretty skilled at this activity.
6. This was an activity that I couldn't do very well. (R)

Appendix E

Study description

Thank you for participating in our study 'Serious gaming for complex decision making'. This study is divided over two days. During the first session (approximately 2.5 hours) of this study, you will play two different scenarios in our serious game and some additional assignments. After each scenario, you will fill in a short survey relating to the scenario. Also, you are going to fill in a questionnaire on demographics.

Appendix F

Debriefing

Thank you for participating in our study 'Serious gaming for complex decision making'. In this message we would like to inform you a little bit more about the purpose of our study. As you may have noticed, while playing the game, some 'rules' in the scenarios and assignments changed. Not only did we want to know how you make your decisions, we also wanted to see how your decision making would be affected by such rule changes. The goal of this study was to see if we can train adaptability through serious gaming with a hint of personalized feedback (depending on the condition you were in).

Thanks again for participating. Your contribution will help us to advance this research! You will receive your reward as soon as possible. After administration is finalized, your responses will be fully anonymized, any identifying data will be destroyed.

As a reminder: All stories were fictitious. Also please do not discuss this study with anybody, so the results will not be confounded should they be willing to participate later on. You are allowed to promote this research, though.

This research is performed in cooperation with TNO, The Netherlands Organisation for Applied Scientific Research.

If you have any questions, feel free to ask those. For more information about this study or questions that arise later, you may contact the experiment coordinator:

Liselotte Eikenhout (email address; phone number).

Appendix G

Normality checks for all measures (Table G1) and an example of Q-Q plots showing deviations (Figure G1).

Table G1

Shapiro-Wilk outcome for ale measures per scenario by condition.

		Shapiro-Wilk (W)			
Measure	Condition	Scenario 0	Scenario 1	Scenario 2	Scenario 3
Proportion TP score	Standardized	,87	,71	,85	,89
	Personalized	,68	,90	,83	,92
Prioritization score	Standardized	,89	,63	,60*	
	Personalized	,90	,64	,56	
Motivation Interest/Enjoyment	Standardized	.92	.92	.95	.97
	Personalized	.97	.91	.96	.93
Motivation Perceived Competence	Standardized	.94	.86	.95	.95
	Personalized	.92	.91	.96	.97
Mental effort learning phase	Standardized	,91	,89	,89	,93
	Personalized	,87	,89	,89	,88
Mental effort test phase	Standardized	,92	,96	,88	,96
	Personalized	,93	,92	,93	,94

Note. When area is marked grey, the measure is not normally distributed (i.e., $p < .05$). * used as an example Q-Q plot showing deviation from normal distribution below (Figure G1).

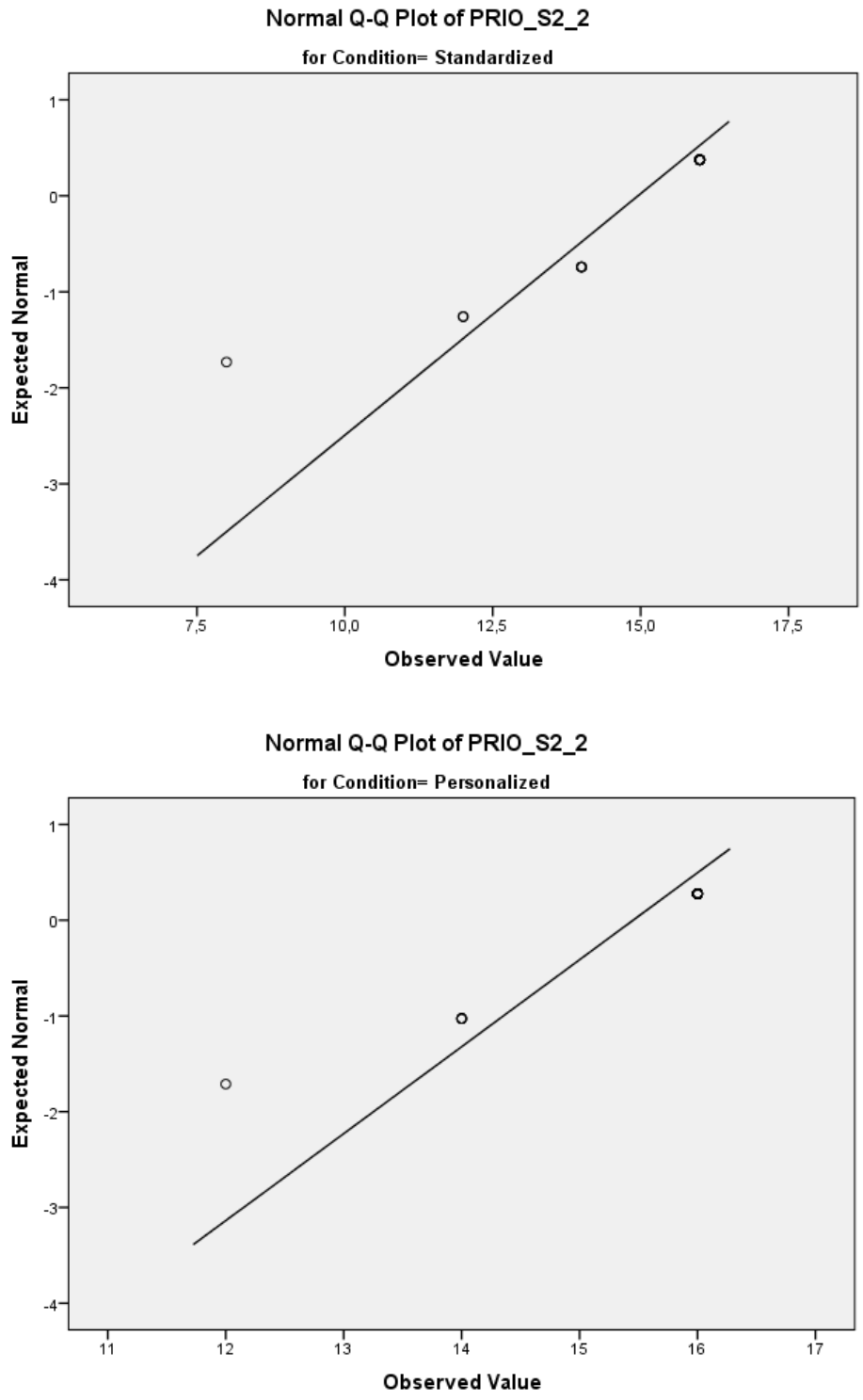

Figure G1. Example Q-Q plots showing deviations from normal for prioritization scores in scenario 2, standardized (above) and personalized (below).