Computers & Education 111 (2017) 101–113

Contents lists available at ScienceDirect

Computers & Education

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

Battling bias: Effects of training and training context

Jackie M. Poos [a, b], Karel van den Bosch, Dr [a, *], Christian P. Janssen, Dr [b]

[a] TNO, Department of Training & Performance Innovations, The Netherlands
[b] Utrecht University, Experimental Psychology and Helmholtz Institute, The Netherlands

A R T I C L E I N F O

Article history:
Received 2 November 2016
Received in revised form 24 March 2017
Accepted 5 April 2017
Available online 10 April 2017

Keywords:
Cognitive bias
Computer games
Critical thinking
Training
Bias mitigation

A B S T R A C T

This study investigates whether cognitive bias in judgment and decision making can be reduced by training, and whether the effects are affected by the nature of the training environment. Theory suggests that biases can be overcome by training in critical reflective thinking. In addition, applied research studies have suggested that game-based training is more effective at reducing bias than conventional forms of training, for example due to the interactive and dynamic nature of video-games. However, earlier studies have not always controlled systematically for the nature of the learning environment between conditions (e.g., providing different content and bias examples for instruction and training). We manipulated in a between-subjects study whether participants received critical-reflective thinking training (yes/no) and in what context they experienced this training (an interactive detective game, or a text-script of the game). Positive effects of training were found. However, the mitigating effects on bias depended upon the type of bias and when the effects were measured (near or far transfer). Surprisingly, the game group performed similar to the text-script group. This suggests that an interactive and dynamic training context (e.g., a game) is not necessarily more effective than non-dynamic contexts (e.g., a text) for bias-mitigation training.

1. Introduction

When humans are faced with unfamiliar and complex situations, a rational and deliberative analysis is not always possible due to, for example, limited information or time pressure. Therefore, people often rely on heuristics to make fast decisions (Kahneman, 2011; Tversky & Kahneman, 1974). Heuristics are mental shortcuts for arriving fast at satisfactory decisions and judgments for example by relying on previous experiences, intuition, rules of thumb, or stereotypes (Kahneman, 2011; Shah & Oppenheimer, 2008; Tversky & Kahneman, 1974).

Although heuristic processing of information can often be very effective, their use can also lead to systematic errors that deviate from logic or rational behavior. Such deviations are referred to as cognitive biases (Dunbar, Miller, Adame, Elizondo, Wilson & Lane, 2014; Kahneman, 2011; Morewedge, Yoon, Scopitelli, Symborksi, Korris & Kassam, 2015; Tversky & Kahneman, 1974). Cognitive biases are believed to be very persistent (Kahneman, 2011), for example due to the evolutionary anchoring of heuristics in the human information processing system (Haselton, Bryant, Wilke, Frederick, Galperina & Frankenhuis, 2009). Moreover, in some situations biases can have dangerous consequences, especially when the safety of

* Corresponding author. TNO, Training & Performance Innovations, PO Box 23, 3769 ZG Soesterberg, The Netherlands.
E-mail addresses: j.m.poos@students.uu.nl (J.M. Poos), karel.vandenbosch@tno.nl (K. van den Bosch), C.P.Janssen@uu.nl (C.P. Janssen).

http://dx.doi.org/10.1016/j.compedu.2017.04.004
0360-1315/

others depends on the judgment and decision-making of professionals such as judges, police, surgeons, and military commanders (Dunbar et al., 2014; Rassin, 2010; Symborski, Barton, Quinn, Morewedge, Kassam & Korris, 2014).

Finding ways to mitigate or eliminate cognitive biases from human reasoning has been a major focus of research in the past few decades (e.g., Evans, 2003; Fischoff, 1981). *Training* people to deliberatively reflect upon the nature of the problem at hand and to critically evaluate the possible solutions, has been suggested to be an effective method for mitigating cognitive biases (e.g., Evans, 2003; Evans, Newstead, Allen, & Pollard, 1994; Fischoff, 1981; Hirt & Markman, 1995; Larrick, Morgan, & Nisbett, 1990; Lilienfeld, Ammirati, & Landfield, 2009; Milkman, Chugh, & Bazerman, 2009; Stanovich & West, 2008). In addition, the application of bias mitigation training in *videogames* has been suggested to be more effective than traditional training methods (e.g., Clegg, Martey, Stromer-Galley, Kenski, Saulnier & Folkestad, 2014; Dunbar et al., 2014; Morewedge et al., 2015; Veinott, Leonard, Lerner Papautsky, Perelman, Stankovic & Lorince, 2013).

Although both training and the use of videogames have been shown to be effective under some circumstances, the reported studies also have their limitations. First, few studies have systematically studied the independent effects of training and the added value of using videogames for training. Instead, the work has focused on comparing game-based training with traditional training methods (i.e., classroom teaching), without ensuring equal training content and teaching philosophy. Second, training studies often focused on bias education (i.e., providing knowledge of the phenomenon 'bias' by teaching definitions) rather than on the skills to recognize the risk of bias in actual situations and on strategies how to prevent bias from occurring (for a discussion on these studies see for example Clegg et al., 2014; Dunbar et al., 2014; Fischoff, 1981; Hirt & Markman, 1995; Larrick et al., 1990; Lilienfeld et al., 2009; Milkman et al., 2009; Morewedge et al., 2015; Veinott et al., 2013). Systematically investigating the factors training and context is important to determine whether videogames as training context provide added value. Third, as noted by others (Hoffman, 2014; Yamnill & McLean, 2001), few studies have tested for transfer of the learned skills to other contexts, outside of the trained context. However, such transfer is essential in order for the training to be of value in everyday life.

The contribution of the current study is to systematically investigate the effects of training and of training environment (i.e., videogames and text) on reducing cognitive biases, while also investigating transfer of learned skills to other contexts. This contributes to the theoretical understanding of bias mitigation and the effectiveness of games for such training. Moreover, this theoretical understanding can be applied to the development and improvement of training scenarios in games. In our study we manipulate two factors: Training ("Training" versus "No training"), and Context ("Game-based" versus "Text-based"). In this way, the isolated effects of both factors as well as their interaction can be identified. Before providing more specific information about the study, we will first discuss relevant theory regarding debiasing training and the effects of presentation context on training in more detail.

1.1. Debiasing training

Although some researchers have expressed pessimism about the feasibility of debiasing (Kahneman, 2011), there is converging evidence that people can often change their behavior for the better. Stanovich & West (2008) identified the following requirements for success: (1) be aware of the rules, procedures and strategies needed to overcome bias, (2) have the ability to detect the need for bias override, and (3) be cognitively capable of decoupling from the bias (Croskerry Singhal & Mamede, 2013).

These requirements have been used, for example, as concepts in a training called *critical thinking* (Cohen & Freeman & Thompson, 1998; Van den Bosch & Helsdingen, 2002). In this training, trainees were taught how to identify (in)consistency and uncertainty, and how to adjust or refine their hypotheses/beliefs about the situation by deliberate testing and evaluation of these hypotheses (Helsdingen, Van den Bosch, Van Gog, & van Merriënboer, 2010). Results showed that this training improved situation awareness and the quality of decision making.

Particular promise comes from critical thinking training methods that teach inferential rules (e.g., consider-the-opposite and consider-an-alternative), often grounded in two-system theories of reasoning (e.g., Lilienfeld et al., 2009; Milkman et al., 2009; Nisbett, Fong, Lehman, & Cheng, 1987). Two-system theories of reasoning assume that people initially make an automatic or intuitive judgment that can be subsequently accepted, corrected, or replaced by more controlled and effortful thinking (Evans, 2003; Kahneman, 2011; Morewedge & Kahneman, 2010; Sloman, 1996). Such effortful thinking is assumed to require time and effort. It is a skill that needs to be learned and trained (Evans, 2003; Evans et al., 1994). People need to learn how to *recognize* situations that demand deliberate, critical and reflective thinking, and to develop the willingness to apply it (Abrami, Bernard, Borokhovski, Wade, Surkes & Tamim, 2008; Evans, 2003; Evans et al., 1994). For example, successful debiasing has been reported from training programs that encourage people to consider and process information that is likely to be underweighted in intuitive judgment (e.g., Hirt & Markman, 1995; Mussweiler, Strack, & Pfeiffer, 2000), or that teach people in statistical reasoning and normative rules of which they may be unaware (e.g., Hirt & Markman, 1995; Larrick et al., 1990).

1.2. Training and training context

Although there is converging theoretical evidence that training itself can be useful, a more open-ended applied question is what the best context is to train in. Recently, studies investigated the assumption that videogames are a more effective medium for providing bias mitigation training compared to more traditional training environments such as classroom teaching or instructional videos (e.g., Clegg et al., 2014; Dunbar et al., 2014; Morewedge et al., 2015; Richey, 2013; Veinott et al., 2013; see for recent review:; Hainey, Connolly, Boyle, Wilson, & Razak, 2016). Noted advantages of videogames are that they can provide realistic, immersive learning environments. Moreover, games support experiential learning by simulating real decision-making scenarios, providing dynamic feedback, providing opportunities to experiment with different actions, and allowing players to observe the consequences (Dunbar et al., 2014; Jarmon, Traphagan, Mayrath, & Trivedi, 2009; Yamnill & McLean, 2001).

Another argument in favor of computer games is that studies using more traditional training tools have rarely found transfer of the learned skills to other contexts (Hoffman, 2014; Yamnill & McLean, 2001). One hypothesis regarding this lack of transfer is that training programs give little attention to how the learned skills can be used in other situations (Yamnill & McLean, 2001). This contrasts with videogames, which theoretically can increase the degree of correspondence between the training context and real-world situations. In effect, this realism of videogames might increase transfer of the learned skills to other contexts.

Several studies have therefore investigated whether videogames are indeed a more effective training environment for mitigating cognitive biases than an instructional video (e.g., Clegg et al., 2014; Dunbar et al., 2014; Lee, Dunbar, Miller, Lane, Jensen & Bessarabova 2016; Morewedge et al., 2015; Veinott et al., 2013). Overall, the results of these studies suggest that game-based training was more effective at cognitive bias mitigation than learning from a video lecture. The game scenes were used as illustrative examples to familiarize the trainees with the phenomenon of cognitive bias, and to teach them the appropriate definitions. Trainees in the control conditions received explanations from the instructor on the same cognitive biases using video. However, the instructor used other content and other examples. It is therefore possible that the reported effects can be attributed to both differences in training content and differences in context: the medium for training delivery (i.e., game or video).

The present study investigates the contribution of the nature of the context to the effects of training by systematically varying the factors training (training versus no-training) and context (game-based versus script-based), while keeping training content equal. Furthermore, the present study attempts to elevate the implementation of principles for effective training in debiasing (as discussed above). Whereas previous work has focused on teaching people the nature and definition of specific cognitive biases, our training focuses on recognizing risky situations and on eliciting critical, reflective thinking.

1.3. Cognitive biases selected for the study

Many dozens of biases have been identified in the literature, each with their own set of determinants. Bias mitigation programs should address the specific principles of the bias (Croskerry, 2015), hence one common approach that affects all biases is not feasible. The focus of the present study is on two specific biases: the *confirmation bias* and the *fundamental attribution error.* The confirmation bias refers to the tendency to selectively search for and consider information that confirm one's beliefs and ignore or overlook information that contradicts it (e.g., Wason, 1968). The fundamental attribution error refers to the tendency of people to attribute the behavior of a person to dispositional influences (e.g., character traits, habits) rather than situational influences (Gilbert, 1998; Jones & Harris, 1967). We use these two biases, as both are well described in the literature. Moreover, in everyday life they can have serious consequences in many professional contexts. For example, police, prosecution, and judges must not rely on first impressions (Fundamental Attribution Error) and remain open to alternative scenarios in which a suspect might be innocent (Confirmation Bias) in order to prevent malfunction of justice (Rassin, 2010).

1.4. The present study

To summarize, previous research on bias mitigation has not systematically investigated the effects of the factors training and context. In addition, previous work has not always measured or found transfer of training to out-of-game contexts. This study addresses both limitations. In our experiment, we created four groups based on the combination of training (training vs no-training) and training context (game-based versus text-based). Half of the participants received training; the other half received no training. Furthermore, half of the participants played a game (an interactive, immersive detective video game); the other half received text (verbal transcriptions of the gameplay scenarios). We tested the effects of training and context on judgment- and decision-making performance by using measurements administered at interruptions during playing the game (game-groups) or during reading the text (text groups). The items of these measurements were linked to events in the game and text, hence they are defined as in-context measurements. In addition, effects were also tested using measurements administered after completion of the game, or reading of the text. The items of these measurements were unrelated to the game and text, hence they are defined as out-context measurements (transfer).

It is expected that *training* (in deliberate, critical and reflective thinking) reduces the likelihood of making biased judgments and decisions compared to practice without training. Moreover, it is also expected that training will be more effective in reducing cognitive bias when it is implemented in a game compared to text, because of the interactive, realistic, and dynamic environment of the game. In the conditions without training we do not expect a difference between the game and text condition. Our additional hypothesis is that training in critical-reflective thinking will produce effects on performance beyond the training context. It is expected that such transfer of training is most likely to occur if training is embedded in a game-based context.

2. Methods

2.1. Participants

Eighty participants (48 women; 32 men) took part on a voluntary basis. All participants were recruited via TNO Soesterberg and Utrecht University. Inclusion criteria were no visual and/or hearing impairments, a good understanding of written English language, and age between 18 and 50 years old ($M = 24.56$, $SD = 5.39$ years), and no familiarity with the videogame Sherlock Holmes: Crimes and Punishment (Paris, France: Frogware, 2014). The study was approved by the internal ethics committee of TNO. Written informed consent was obtained from all participants. Participants were reimbursed with 15 euros per hour.

2.2. Design

We used a 2 (training: yes, no) x 2 (context: game or text) between subjects design, which resulted in four groups. Each group had 20 participants. We balanced groups in terms of their cognitive flexibility, as cognitive flexibility might affect a participant's tendency to make biased judgments and decisions and therefore their potential benefit from critical-reflective thinking training (see also Martin & Rubin, 1995). Before participants arrived in the lab, we measured cognitive flexibility using the test by Martin and Rubin (1995). Based on test performance, we assigned participants to the four groups, so as to balance the cognitive flexibility scores across groups. In general, the majority of participants had a medium high or high score, perhaps because most were students (scores between 39 and 72 out of 72). Only one participants scored medium-low; none of them had a score in the low category. Participants from the different categories were equally divided over the experimental conditions. All groups received in-context measurements, and out-context measurements.

2.3. Materials

2.3.1. Game context

Half of the participants experienced a game context. In this condition, participants played Sherlock Holmes: Crimes and Punishments (Paris, France: Frogware, 2014). We embedded our bias-mitigation intervention in this modern commercial off-the-shelf video game because the game has been rated positively and to convincingly address the skills of interest: "*In addition to finding the killer or killers, you also rationalize why the crime was committed and select between condemnation or absolution.*" (Todd, 2014). Within the game, participants were presented with two murder cases. In the first murder case ('The fate of Black Peter'), participants played a point-and-click detective adventure in which they controlled the third-person view game character Sherlockz Holmes. Participants played cooperatively with two non-playable characters: Dr. Watson and Inspector Lestrade. The game required participants to interrogate witnesses and search crime scenes for evidence in order to find and arrest a murderer. For each murder case four possible suspects were presented during the gameplay. After the game was finished, the game informed participants whether or not they had arrested the actual murderer.

During piloting we found that playing the game without any help took a very long time. Therefore, a help system was designed to speed up gameplay. This system provided the participant with additional procedural instructions and hints on how to proceed in the game (e.g., 'walk down the path'). Even with the help system, playing two murder cases proved to take long (i.e., around 6–7 h). Therefore, the second (different) murder case ('The Kew Gardens drama') was presented as a movie of the gameplay. Although a movie clip preserves the dynamic nature and visual richness of a game, it does not contain the feature of interactivity which is an attractive feature of a game.

2.3.2. Text context

Participants who did not play the game, experienced the text context (half of participants). In this condition they read stories of the murder cases on paper. The stories were written up by the experimenter based on the game-play. The text had the same content as the game-context. For example, it contained all the dialogues and descriptions of the environment where Sherlock walked. The big contrast with the game context was that the script did not offer the visual richness and interactivity of the game. In addition, in contrast to participants in the gaming conditions, participants in the text-based conditions could not decide which suspect to arrest. Instead, they read how Sherlock determined who was guilty at the end of the stories.

2.3.3. Training

Participants in the training conditions (for both the game-based and text-based context) received critical-reflective thinking training during the first murder case only (i.e., all participants played or read the second murder case without training). During the training case, participants were told that they would receive feedback from an experienced detective, Hercule Poirot, to make them become better detectives. The training instructions were administered immediately following each in-context measurement for the biases: Confirmation Bias (2 training moments) or Fundamental Attribution Bias (4 training moments), explained in more detail later). The training instruction consisted of the following elements in chronological order:

1. Participants were presented with a conjecture of Sherlock that might be the result of a cognitive bias. The mechanism underlying the possible bias was explained, without using terms as "bias" or "error". In particular, the risk of committing a fundamental attribution error was called "impression-thinking". The risk of committing confirmation bias was referred to as "confirmation-thinking". The explanations were the most elaborate after the first CRT training moment. In subsequent training moments, the explanations were shorter. We did not use the terminology bias, as in contrast to earlier work (e.g., Clegg et al., 2014; Dunbar et al., 2014; Symborski et al., 2014; Veinott et al., 2013) our method focused on 'learning by doing' instead of learning definitions.
2. An example of how the bias may manifest itself in a real-world situation was given.
3. A strategy was explained regarding how to recognize situations that are likely to evoke bias and how to prevent actually making a biased assessment (Cohen et al., 1998; Helsdingen et al., 2010; Van den Bosch & Helsdingen, 2002). For countering impression-thinking (the fundamental attribution error) this involved (in summary): "check whether your explanation of someone's behavior is based upon your impression of that person. Then, think of another explanation based upon the circumstances that may also account for someone's behavior. Reconsider both explanations and choose the most plausible". For countering confirmation-thinking (confirmation bias), the strategy involved: "identify a critical assumption underlying your assessment of the situation; think of which alternative circumstances may prove the assumption to be wrong; then check for evidence or plausibility of such circumstances. Finally, reconsider the original and the criticized situation assessment, and choose the most plausible".
4. A worked-out example was presented showing how to use these approaches in an example situation from the game/text-narrative.
5. Then subjects were given an assignment, with a situation taken from the game/text-narrative, wherein participants had to apply the instructed approach

After completing a training instruction session, participants continued playing the game or reading the story until the next in-context measurement, or until completion of the murder case.

2.4. Dependent measures

2.4.1. Research Question 1: In-context measurements

The game play and reading of the stories were interrupted fourteen times for in-context measurements. The measurements were developed based on the narrative of the game/stories. In total, eight Fundamental Attribution Error (FAE) and six confirmation Bias (CB) measurements were administered. Measures collected during the first murder case were not used for analysis, as in that part of the study the participants in the training-conditions received instruction and practiced on bias detection and bias prevention (see above). These in-between instruction and practice sessions are likely to affect the measurements. Only the measurements on the second murder case were used. The order of measurements in murder case 2 was: FAE, FAE, FAE, CB, CB, FAE, CB, CB. During the second murder case, no instruction and practice was given.

2.4.1.1. Fundamental attribution error measurements. During the game/stories several suspects had to be interrogated. After a suspect had been interrogated, a multiple choice question was administered which measured the fundamental attribution error (see Appendix A for an example). Participants were presented with a hypothesis of Sherlock Holmes at that moment in the game/story (e.g., "John is guilty"). Six pieces of information pertaining to the hypothesis were also given: three presented dispositional information about the suspect, and three presented situational influences. The pieces of information were categorized by the experimenter and independently validated by two colleagues who are experienced researchers in the domain of decision making, and who were not involved in the present study. They identified these unequivocally as intended.Participants were asked to select the three pieces of information they considered as most important to gain more certainty about the hypothesis. A participant's tendency to commit a fundamental attribution error was determined by calculating the mean number of dispositional responses made during the second murder case.

2.4.1.2. Confirmation bias measurements. Once multiple pieces of evidence in the game and stories pointed towards a particular suspect, a Confirmation Bias measurement was administered. Participants were presented with a hypothesis of Sherlock about the guilt or innocence of the suspect, including a motive, a murder weapon, and time and place of the murder (see Appendix A for an example). Again, six pieces of evidence were given. Three of these were in conjunction with the hypothesis; three others were in conflict with the hypothesis. Again, this categorization was checked and confirmed by two colleague researchers prior to the experiment proper. Participants were asked to select three pieces of evidence they considered as most important to gain more certainty about the hypothesis. A participant's tendency to commit a confirmation bias was determined by calculating the mean number of responses involving confirmatory evidence.

2.4.2. Research Question 2: Out-context measurements

2.4.2.1. Fundamental attribution error measurement. The Fundamental Attribution Error test used by Dunbar et al. (2014) was used. The test was translated into Dutch, which necessarily involved some slight edits in the exact phrasing of questions. The test incorporated eight items in total (see Appendix B for an example). Four were included as pretest, four as posttest. Each item first provides a one-paragraph narrative of a character that experiences a series of either fortunate or unfortunate events. Then six different explanations were given for the events. Three of these used dispositional information about the character to account for the events, three other explanations referred to situational influences. Participants were asked to choose three explanations that best explained the series of events. A participant's tendency to commit a fundamental attribution error was determined by calculating the number of dispositional responses for each item.

2.4.2.2. Confirmation bias measurement. Two out-context measures of Confirmation bias were used: the NewCB and CBAM test. First, the NewCB of Dunbar et al. (2014) was translated into Dutch. Seven of the 14 items were administered at the prestest, the other seven at the posttest (see Appendix B for an example). Each item first provided a one-paragraph narrative about an event that may have several causes, followed by a hypothesis about how it happened. Then, three pieces of evidence were given. One was in conjunction with the hypothesis, the other two were in conflict with the hypothesis. Participants were asked to select the piece of evidence they considered as most important to gain more certainty about the hypothesis. A participant's tendency to commit a confirmation bias was determined by calculating the proportion of responses with conjunction evidence.

The second test was the CBAM test, also developed and validated by Dunbar et al. (2014), and was translated into Dutch. The CBAM test consisted of six items. Three items were used in the pretest, three in the posttest (see Appendix B for an example). Each item presents a one-paragraph narrative followed by a hypothesis about how it happened. Then, four pieces of evidence were given. Two were in conjunction with the hypothesis, the other two were in conflict with the hypothesis. Participants were asked to select the two pieces of evidence they considered as most important to gain more certainty about the hypothesis. A participant's tendency to commit a confirmation bias was determined by calculating the number of responses with conjunction-evidence. In order to create one encompassing out-context measurement of the tendency to commit confirmation bias, mean scores on the NewCB test and the CBAM test were used to calculate a pretest-CB and a posttest CB score.

2.4.3. Additional measures

2.4.3.1. Demographic data. Demographic data such as age, gender and experience with videogames was collected in a pre-lab survey.

2.4.3.2. Cognitive flexibility. We used Martin & Rubin (1995) 10-item self-assessment cognitive flexibility scale in the pre-lab survey to assign participants equally to conditions (see design).

2.4.3.3. Cognitive Reflection Test. To measure participants' inclination to respond intuitively to a problem (versus reflecting on it), we used the three-item Cognitive Reflection Test (Frederick, 2005) in the pretest, and the four extra items of the expanded version (Toplak, West, & Stanovich, 2014) during the posttest. Data on the second extra item were deleted from analyses due to an incorrect translation. For both the pretest and posttest, the proportion correct responses was calculated.

2.4.3.4. Immersion questionnaire. To measure the immersiveness that participants experienced when playing the game or reading the story, the 30-item immersion questionnaire of Jennet, Cox, Cairns, Dhoparee, Epps & Tijs (2008) was used. Two Dutch versions of the questionnaire were created: one tuned to the game-based conditions, the other tuned to the text-based conditions. The immersion questionnaire was administered immediately following the first murder case. Each item in the questionnaire represented a statement which participants had to rate from 1 (not at all) to 5 (very much). Due to the difference in presentation, some questions of the questionnaire did not apply to the text condition (e.g., questions about graphics). We therefore did not use the raw score which had a higher maximum score in the game condition (30 questions, maximum 150 points) compared to the text condition (28 questions, 140 points). Instead, we calculated how high the score was as a proportion of the maximum obtainable score per condition.

2.5. Procedure

Participants were recruited via e-mail, social media, and an information letter. After expressing interest, participants filled out an online survey for gathering demographics (to check eligibility), cognitive flexibility, and cognitive reflection measures. Based on the cognitive flexibility measure, participants were assigned to one of four conditions (see design).

The experiment itself was conducted in a quiet room. Participants sat in front of either one laptop (text-based conditions) or two laptops (game-based conditions). After informed consent was signed, an experimenter administered pretest measures, followed by the first murder case with the in-context measurements and the training (if applicable). After completion of murder case one, the immersion questionnaire was administered, followed by a short break. Then, participants either watched a video of a gameplay (game-context), or read a second story (text-context) of murder case two, and received the in-context measurements. After completion of murder case two, the post-test was administered.

In total, the game-based conditions took approximately 4 h to complete. Text-based conditions took approximately 2 h to complete.

3. Results

3.1. Prerequisite tests

An assumption in game training is that games provide a more immersive experience compared to other training environments. A one-way ANOVA revealed no significant effect of Context on the immersion scores, ($F(1,78) = 0.43$, $p = 0.51$). Participants in game-based conditions ($M = 0.60$, $SD = 0.04$; raw score: 90.18 points out of maximum of 150 points) did not have higher immersion scores than participants in text-based conditions ($M = 0.59$, $SD = 0.05$; raw score: 83.1 out of 140).

3.2. Research Question 1: The effects of training and context on the in-context measurements

3.2.1. Effects on FAE in-context measurements

A 2 (Context) x 2 (Training) two-way ANOVA was performed on the FAE-score on the in-context measurements of murder case two. Analysis revealed a significant main effect of Training ($F(1,73) = 14.12$, $p < 0.01$, $\eta^2 = 0.16$). As Fig. 1 shows, participants in the Training conditions ($M = 0.10$, $SD = 0.11$) gave fewer FAE-biased judgments and decisions than participants in the No-Training conditions ($M = 0.22$, $SD = 0.16$). We found no significant main effect of Context ($F(1,73) = 0.80$, $p = 0.37$), nor an interaction effect between Training and Context ($F(1,73) = 0.64$, $p = 0.43$).

3.2.2. Effects on CB in-context measurements

A 2 (Context) x 2 (Training) ANOVA was performed on CB-score on the in-context measurements of murder case two. Analysis revealed no significant main effects of Training ($F(1,73) = 1.94$, $p = 0.17$), and no effect of Context ($F(1,73) = 1.01$, $p = 0.32$). In addition, no significant interaction effect between Training and Context was found ($F(1,73) = 0.16$, $p = 0.69$, $\eta^2 < 0.01$).

3.3. Research Question 2: The effects of training and context on the out-context measurements

3.3.1. Effects on FAE out-context measurements

A 2 (Test Moment) x 2 (Context) x 2 (Training) mixed ANOVA was performed on FAE-scores on the out-context measurements. Three participants (one in the Text-Based No-Training condition and two in the Text-Based Training condition) did

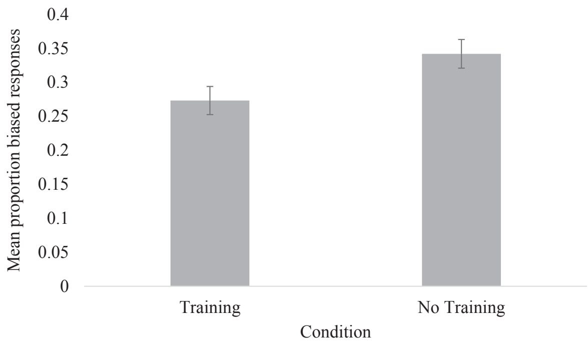

Fig. 1. Mean proportion biased responses of the Training versus No-Training groups on the Fundamental Attribution Error in-context measurements of murder case two. Error bars show standard errors.

not complete the CB questionnaire due to an administration error and were left out of the analysis. No significant main effect was found for Test Moment (F (1,76) = 2.04, p = 0.16). Moreover, no significant main effects were found for Training (F (1,76) = 0.17, p = 0.68), or for Context (F (1,76) = 0.14, p = 0.71). In addition, no interaction effects were found (all $ps > 0.05$).

3.3.2. *Effects on CB out-context measurements*

A 2 (Test Moment) x 2 (Context) x 2 (Training) mixed ANOVA was performed on the CB-scores on the out-context measurements. One participant (in the Game-Based No-Training condition) did not complete the CB questionnaire due to an administration error. The data of the remaining 79 were used in the analysis. The results revealed a significant main effect of Test Moment (F (1, 75) = 18.43, $p < 0.01$, $\eta^2 = 0.20$), revealing that participants gave fewer CB responses on the posttest compared to the pretest. In addition, a significant effect of Training was found (F (1, 75) = 16.51, $p < 0.01$, $\eta^2 = 0.18$). A significant interaction effect between Test Moment and Training was also found (F (1, 75) = 21.34, $p < 0.01$, $\eta^2 = 0.22$). This effect is illustrated in Fig. 2. Within-group analysis revealed that participants who received training reduced their proportion of CB-biased responses significantly from pre-to posttest (from 0.62 to 0.45, F (1,77) = 41,03, $p < 0.01$). In contrast, participants that received no training did not change on this measure from pre-to posttest (0.68 on both tests). No main effect of Context was found (F (1, 75) = 0.82, p = 0.37), nor an interaction effect between Training and Context (F (1, 75) = 2.95, p = 0.09).

3.3.3. *Effects on the Cognitive Reflection Test*

A 2 (Test Moment) x 2 (Context) x 2 (Training) mixed ANOVA was performed on the proportion correct responses on the Cognitive Reflection Test. No significant main effects were found for Test Moment, (F (1,76) = 1.86, p = 0.18). In addition, no significant main effects were found for Training, (F (1,76) = 0.06, p = 0.80), or for Context, (F (1,76) = 0.14, p = 0.71). All two-way interactions and the three way-interaction were not significant, all $ps > 0.05$. Thus, all conditions performed similarly on the pretest and the posttest.

4. General discussion

4.1. *Game contexts do not necessarily offer an advantage*

This study investigates systematically whether bias mitigation improves due to *training* (is a training provided or not?) and whether the *context* of training is important (use of a game, compared to a text). This was investigated for two biases: the Fundamental Attribution Error and the Confirmation Bias. Performance on both biases was measured on in-context measurements (Research Question 1) and out-context measurements (Research Question 2). Our findings show that embedding training in the context of a game provides no additional performance benefits, when compared to a training using text-based narratives. Groups performed equal with respect to committing or avoiding bias, on both the in-context and the out-context measurements.

This result contrasts with earlier work that reported superior performance of training in the context of a game compared to a video lecture (Clegg et al., 2014; Dunbar et al., 2014; Richey, 2013; Symborski et al., 2014; Veinott et al., 2013). However, as we argued in the introduction, this previous work did not control for the content of both the contexts: all these studies compared training in a video game with watching a video of other content. Therefore, this work could not separate the effects of content from the effects of context. In our study, the content of both conditions was the same (the narrative of a game), but the context was different (playing the game or reading a script).

Two other explanations for the results of the previous studies might be that the video context and the game context differed in (1) how much they captured the participants' attention, or (2) how intrinsically motivating the task was. Capturing the attention of a trainee is one of the most noteworthy challenges of modern training programs, as it requires an activity to be “fun, interesting, captivating, appealing and intrinsically motivating” (Richey, 2013). Similarly, it has been argued that intrinsic motivation promotes learning and in particular that games are effective because they are intrinsically motivating to

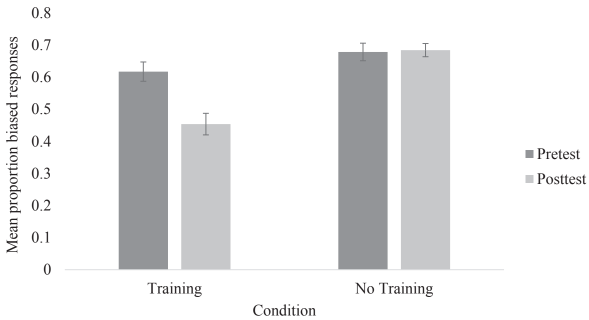

Fig. 2. Mean proportion of Confirmation Bias responses for the Training and the No-Training group on the out-context measurement. Error bars show standard errors.

players (McGonigal, 2011). A game might be more appealing on both capturing a player's attention and increasing their intrinsic motivation, compared to a video.

Any limitations with respect to motivation and attractiveness of the context in the comparison group may have been less severe as in previous studies by others, as in our study the text-based context offered the same narrative as the game, and the use of a narrative has been found to increase learning and engagement and to provide motivating learning scenarios (Hinyard & Kreuter, 2007; Hinyard, Green, Cappella, Slater, Wise & Storey, 2007; Detorri & Paiva, 2009; Rowe, McQuiggan, & Lester, 2007). This is also reflected in our immersion scores: participants rated the game equally immersive as the script.

The equal immersion scores between the text and game condition suggest that a narrative is sufficient to engage students in immersive learning, and that adding interactive game-play elements has no added value. Even though immersion scores were not high, this does not affect the claim that game-based training contexts are not necessary for training, good narrative scripts might also be sufficient. Most of the earlier mentioned studies into game-based learning of bias prevention in decision making (Clegg et al., 2014; Dunbar et al., 2014; Lee et al., 2016; Morewedge et al., 2015; Veinott et al., 2013) did not include or used different measures of immersion. Therefore, our scores cannot be compared to those studies.

Studies that did investigate immersion in games, did not study immersion in a debiasing context, and are therefore also hard to compare to our results. Such studies found, on average, higher immersion scores (e.g., Denisova & Cairns, 2015a, 2015b). Several factors may explain the over-all lower immersion scores in our study context. One factor is that participants had to pause their game-play at predetermined moments to engage in training activities and in-context measurements. These interruptions may have affected their immersion ratings. Another factor is that the original immersion questionnaire was translated into Dutch, possibly introducing subtle differences in the connotations of the questions. Yet another factor is that the original questionnaire had to be adjusted for use in this study (some questions were not suitable for testing immersion in text-based learning environments). These circumstances make it necessary to exercise caution in comparing our immersion scores to that of other studies.

4.2. Training effects

A remarkable outcome of our study is that the positive effects of training in critical-reflective thinking on reducing the two investigated types of bias were demonstrated in different phases. The tendency to commit Fundamental Attribution Bias was only reduced by training on the in-context measurements (see Fig. 1), whereas the tendency to commit Confirmation Bias was not reduced on in-context measurements, but only on out-context measurements (see Fig. 2). We will discuss the two diverging patterns successively.

The positive effects of training on reducing Fundamental Attribution Bias, as observed in the training condition, did not transfer to the independent posttests. Failing to apply skills to other situations than where they were acquired is, unfortunately, not uncommon. Hoffman (2014) have described several reasons why transfer does not happen when it is easy to imagine that it should. In general, skills and knowledge are considered to be task, domain, and context specific. This makes it possible that two tasks can be manifestly similar, and yet carry some apparently minor differences that could lead to no or even negative transfer ("transfer surprises", see Hoffman, 2014; Lewandowsky, Kalish, & Ngang, 2002).

The tasks that we used to measure the participants' inclination to commit fundamental attribution bias, both for in-context and out-context measurements, were similar in some respects (i.e., both text-based multiple-choice questions). However, the differed in two respects. First, in the context: rich contextual narratives in the in-context measurements, versus concise and global short stories in the out-context measurements. Second, they differed in the phrasing of the question. In-context measurements required subjects to select the three pieces of information (from a series of six) that they considered as most important to determine the guilt or innocence of particular character. All presented information items were deeply connected to the narrative, requiring subjects to judge very carefully. The out-context measurements on the other hand, presented eight one-paragraph descriptions of a day in the life of a fictitious character that either passed successfully or disastrous. Subjects were asked to select the three reasons (from a series of six alternatives) they considered the most important for explaining the events happening to the character. Many of these alternatives had a general nature and were re-used in every item (e.g. 'personality'; 'skills'; behavior of character's partner(s)). The similarity in descriptions of the 'day-in-the-life-of'-items, and the re-occurring response-alternatives may have induced a more superficial analysis than on the in-context measurements.

These differences may have been the reason for not finding transfer-of-training effects with respect to the fundamental attribution bias. As the need for transfer from training to other settings is a self-evident requirement (Clegg et al., 2014), this result makes clear that the training process needs to be improved in order to be of use for practical applications. Whether improvements should involve adaptations of the training process, of the training volume, or both is a matter of further research.

The reversed pattern of results was found for the Confirmation Bias: these were not reduced on in-context measurements, but were reduced from pre-to post-tests on out-context measurements. This demonstrates that participants were able to apply the skills they learned during training in another context, but not necessarily in the game context. There are two potential explanations for this pattern of results. First, it could be that the game and story induced such robust confirmatory thinking that participants were unable to challenge their intuitive beliefs or hypotheses. Second, the elaborate, diverse and rich narratives and rich information that is provided in the game and the written narrative, may have made it too difficult for participants to inhibit the inclination towards confirmatory thinking during the in-context measurements. In contrast, the out-context measurements provided shorter, simpler, printed situation descriptions in which it could have been easier to apply the learned approach from training.

Beyond investigating whether training affected participants' tendency to commit bias, we also measured whether the training as a whole induced participants to employ a more reflective – instead of an intuitive – approach when facing new problems. We used the Cognitive Reflection Test (Toplak et al., 2014) to measure this. Previous research has shown that performance on the Cognitive Reflection Test has predictive value for performance on heuristics-and-biases tasks (e.g., Hoppe & Kusterer, 2011; Oechssler, Roider & Schmitz, 2009; Toplak, West & Stanovich, 2014). It is therefore argued that the test is appropriate for measuring people's inclination to incorrectly use heuristics because the test presents a short series of problems that prime a strong alternative -but incorrect-response. It is assumed that the participant must inhibit this prompted alternative and conduct deliberate and critical reflection on the problem to come up with the correct answer.

In the present study, we did not find any effects of training in critical-reflective thinking on the Cognitive Reflection Test metric. One conclusion may be that the training did not bring about a fundamental change in the way participants approach new and unfamiliar problems. There are, however, also indications that the Cognitive Reflection Test may not be a suitable instrument to measure (changes in) one's ability to reflect upon regularities, relationships and inferences in uncertain and inconsistent environments. Instead, it is argued that the test primarily calls upon a person's numeric ability (Sinayev & Peters, 2015). If this is the case, then the Cognitive Reflection Test was not the proper task to demonstrate effects of training since our critical-reflective thinking training focuses on cognitive, reflective abilities and *not* on numeric ability.

4.3. Implications

An important implication of this work is that it shows that games are not a panacea for bias reduction training. When the training environment is controlled for the content of training (as done here), then other media can provide comparable results. To advance theory, more research is needed to understand what aspects of a learning environment provide the best context for bias-mitigation training. Our work suggests that having an engaging narrative is one contributing factor, which might be abundant to achieve training. This supports the findings of others that also argue that narrative is also one important contributing factor (cf. Hinyard & Kreuter, 2007; Hinyard et al., 2007; Detorri & Paiva, 2009; Rowe et al., 2007).

An implication for practice is that the development of a high immersive game to achieve effective bias mitigation training might not always be needed. This is particularly relevant for organizations that have scarce resources to spend on developing training materials. Instead of spending these resources on a potentially expensive game, they can dedicate these resources to other aspects that can make training successful.

4.4. Limitations and future directions

We used the publicly available videogame "Sherlock Holmes" in this study as pilot tests of various games showed that this game provided an immersive experience. However, a limitation of using a commercial product for research purposes is that it is impossible to fully modify the game as desired. As a consequence, playing the game took a very long time, as game scenarios were relatively long. It was therefore not feasible to let participants play both murder cases. The training instructions and exercises were therefore administered in the context of the interactive game (murder case one), and measurements were collected during a movie of the gameplay (murder case 2). Effects of training are therefore established during the phase that participants interactively play the game. So outcomes of presentation context on training (script-based training has equal effects as game-based training) cannot be attributed to the fact that murder case 2 was presented as a movie of the game, rather than as an interactive game. Taken together, the often assumed added value of games to training is not yet an open-and-shut case. Further research into This is needed. Preferably, the context for administering training and for testing the effects of training should have a more constant implementation than used in this study. This could be achieved by using another suitable but shorter game, or by developing a dedicated game, specifically designed and tunable to the study purposes.

Furthermore, although the present study provides evidence that critical reflective thinking training successfully reduced the inclination to make biased responses, it did not consistently do so for in-context and out-context measurements and for both the trained biases. More work is needed to find out what the cognitive and situational factors are that determine the effects of training. Moreover, as the spectrum of human biases is varied, more work is needed to investigate how effective training is for these various other biases.

5. Conclusion

Our work shows that critical reflective thinking training reduces the inclination to make biased responses, and there is evidence that for one bias this effect extends beyond the training tasks. The positive effects are achieved irrespective of training context: games are not better than narrative texts.

Acknowledgments

This research has been funded by the Netherlands Ministry of Defense (grant number: V1522). The funding organization had no involvement in the nature or design of this study.

Appendix A. Examples of in-context measurements

Example of question measuring proneness to fundamental attribution error

Sherlock Holmes just spoke to Judith Carey, the wife of the victim, and has the feeling that she is probably innocent and has nothing to do with the murder on her husband. Sherlock searches the police archives for information on Judith Carey.

Below is a list of information about Judith Carey retrieved from the police archives. Imagine that you are Sherlock Holmes and you want to investigate the assumption of her innocence. Which pieces of information about Judith Carey are most important to you? Choose three (no less, no more).

a. Judith Carey and her husband were not particularly wealthy or prosperous. Judith Carey will not be left an inheritance after her husband's death.
b. Judith Carey was completely dependent upon Peter Carey. Financially, as well as in other matters, he took care of her. It will be difficult for Judith to support herself after her husband's death.
c. Judith Carey is a very religious and god-fearing woman whose norms and values are very important to her. Because of this, she would never commit a murder.
d. Judith Carey is a small, not very strong woman. It would be very difficult for her to murder a strong man such as Peter Carey.
e. Judith Carey is a kind and loving person, despite her embittered husband's abuse.
f. Judith Carey was a hopeful person. She expected Peter Carey to return to the man he once was.

(alternatives a, b, & d present situational information; alternatives c, e, & f present dispositional information).

Example of question measuring proneness to confirmation bias

"Liam Hurtley murdered Peter Carey out of revenge. Liam Hurtley is a veteran who became the gardener at Woodman's Lee after retiring from service. He fell in love with Peter Carey's wife, Judith. Liam Hurtley could not stand seeing Peter Carey physically abusing his wife and at a certain point he killed Peter Carey."

This is a hypothesis that Sherlock Holmes formed about the murder on Peter Carey. He checks his notes on available evidence. Imagine that you are Sherlock Holmes and you want to investigate the hypothesis. Which pieces of evidence are most important to you? Choose three (no more, no less).

a. Liam Hurtley is the gardener at Woodman's Lee. When you asked Judith Carey whether there was a gardener at Woodman's Lee, she lied about this.
b. The murder weapon is a harpoon. Peter Carey's body was impaled to the wall with this harpoon, requiring immense strength and agility in throwing harpoons.
c. A tobacco pouch with the initials P.C. were found on the table in Peter Carey's cabin. However, according to Judith Carey her husband stopped smoking a long time ago.
d. On the table in Peter Carey's cabin, two glasses were found with rum of a brand that is popular among sailors. This suggests that Peter Carey was drinking with someone else on the night of the murder, probably with somebody he knew.
e. A stained letter was found suggesting that Liam Hurtley hid something at Woodman's Lee. The letter states the following: "I did as you asked and hid them well … but I beg you to reconsider. Remember our vows! I have done nothing dishonorable for which you … me."
f. From several hidden letters, it became apparent that Judith Carey and Liam Hurtley had romantic feelings for each other. However, at this point it is unclear how far this relationship went.

(alternatives a, e, and f present information in conjunction with the hypothesis; alternatives b, c, and d are not consistent with the hypothesis).

Appendix B. Examples of out-context measurements

Example of a question of the NewCB test

Your colleague, Diane, took the bicycle to go to work. There are several explanations for her travelling by bike. You think she came cycling because she likes to exercise. How would you check whether your assumption is right? Tick one action.

a) Find out whether Diane thinks physical exercise is important
b) Find out whether Diane thinks there are too few parking spots at work
c) Find out whether Diane's car broke down

(Alternative a is an action that can confirm the hypothesis, the other alternatives check for other possibilities).

Example of a question of the CBAM test

You come home one night and you notice that the lamp in the living room is broken. You suspect that one of your roommates did it, because he is rather clumsy, especially when he has been drinking. Besides he was probably in the living room earlier this night. You may choose two questions to investigate your suspicion. Which two do you choose? (no more, no less).

a. Are there reasons to believe that someone else (than your roommate) broke the lamp?
b. Has your roommate been drinking that day?
c. Is there any evidence that your roommate was not home earlier that night?
d. Has anybody seen your roommate home that night?
Alternatives b and d are questions that attempts to verify the hypothesis, alternatives a and c are questions to disprove the hypothesis)

Example of a question of the fundamental attribution error test

Ron's Bad Day

Ron was a bit late to work on Monday morning because there was a bad traffic jam. When he arrived at work he didn't feel very well and got a slow start on his work for the day. He made it to his 10:00 a.m. staff meeting, but because that meeting ran 10 min late, he arrived late at his 11:30 a.m. sales team meeting. At that meeting Ron and three of his coworkers were supposed to present a new production line to the Director, but the presentation wasn't finished, and the laptop projector wasn't functioning properly. At 2:30 p.m. when Ron was back at his desk, three clients called and cancelled their long-standing monthly supply contracts with Ron's firm. At 4:00 p.m. Ron was called in to the director's office and fired.

Choose three (no more, no less) of the following options that you think are most likely to have caused what happened to Ron in the above scenario.

a. The traffic jam
b. Ron's skills
c. Ron's personality
d. Ron's capacities
e. Ron's clients
f. Ron's director

(alternatives a, e, & f present situational information; alternatives b, c, & d present dispositional information).

References

Abrami, P. C., Bernard, R. M., Borokhovski, E., Wade, A., Surkes, M. A., Tamim, R., et al. (2008). Instructional interventions affecting critical thinking skills and dispositions: A stage 1 meta-analysis. *Review of Educational Research, 78*(4), 1102–1134.

Clegg, B. A., Martey, R. M., Stromer-Galley, J., Kenski, K., Saulnier, T., Folkestad, J. E., et al. (2014). Game-based training to mitigate three forms of cognitive bias. In *Paper presented at interservice/industry training, simulation and education conference, Orlando* (pp. 1–12).

Cohen, M. S., Freeman, J. T., & Thompson, B. (1998). Critical thinking skills in tactical decision-making: A model and a training strategy. In J. A. Cannon-Bowers, & E. Salas (Eds.), *Making decisions under stress: Implications for individual and team training* (pp. 155–190). Washington, DC: APA.

Croskerry, P. (2015). When I say… cognitive debiasing. *Medical Education, 49*(7), 656–657.

Croskerry, P., Singhal, G., & Mamede, S. (2013). Cognitive debiasing 1: Origins of bias and theory of debiasing. *British Medical Journal: Quality & Safety, 22*(Suppl 2), ii58–ii64.

Denisova, A., & Cairns, P. (2015b). Adaptation in digital games: The effect of challenge adjustment on player performance and experience. In *Proceedings of the 2015 annual symposium on computer-human interaction in play, London, UK* (pp. 97–101). New York: ACM Publications.

Denisova, A., & Cairns, P. (2015a). First person vs. third person perspective in digital games: Do player preferences affect immersion?. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems, Seoul, Republic of Korea* (pp. 145–148). New York: ACM Publications.

Detorri, G., & Paiva, A. (2009). Narrative learning in technology-enhanced environments. In N. Balacheff, S. Ludvigsen, T. de Jong, A. Lazonder, & S. Barnes (Eds.), *Technology-enhanced learning* (pp. 55–69). Dordrecht, the Netherlands: Springer.

Dunbar, N. E., Miller, C. H., Adame, B. J., Elizondo, J., Wilson, S. N., Lane, B. L., et al. (2014). Implicit and explicit training in the mitigation of cognitive bias through the use of a serious game. *Computers in Human Behavior, 37*, 307–318.

Evans, J. S. B. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences, 7*(10), 454–459.

Evans, J. S. B., Newstead, S. E., Allen, J. L., & Pollard, P. (1994). Debiasing by instruction: The case of belief bias. *European Journal of Cognitive Psychology, 6*(3), 263–285.

Fischoff, B. (1981). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty. Heuristics and biases* (pp. 422–444). New York, NY: Cambridge University Press.
Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives, 19*(4), 25–42.
Frogware. (2014). *Sherlock Holmes: Crimes and Punishments*. Paris, France: Focus Home Interactive.
Gilbert, D. T. (1998). In S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology*. New York, NY: McGraw-Hill.
Hainey, T., Connolly, T. M., Boyle, E. A., Wilson, A., & Razak, A. (2016). A systematic literature review of games-based learning empirical evidence in primary education. *Computers & Education, 102*, 202–223.
Haselton, M. G., Bryant, G. A., Wilke, A., Frederick, D. A., Galperin, A., Frankenhuis, W. E., et al. (2009). Adaptive rationality: An evolutionary perspective on cognitive bias. *Social Cognition, 27*(5), 733.
Helsdingen, A. S., Van den Bosch, K., Van Gog, T., & van Merriënboer, J. J. (2010). The effects of critical thinking instruction on training complex decision making. *Human Factors, 52*(4), 537–545.
Hinyard, M. W., Green, M. C., Cappella, J. N., Slater, M. D., Wise, M. E., Storey, D., et al. (2007). Narrative communication in cancer prevention and control: A framework to guide research and application. *Annuals of Behavioral Medicine, 33*(3), 221–235.
Hinyard, L. J., & Kreuter, M. W. (2007). Using narrative communication as a tool for health behavior change: A conceptual, theoretical, and empirical overview. *Health Education & Behavior, 34*, 777–792.
Hirt, E. R., & Markman, K. D. (1995). Multiple explanation: A consider-an-alternative strategy for debiasing judgments. *Journal of Personality and Social Psychology, 69*(6), 1069.
Hoffman, R. R. (2014). *The psychology of expertise: Cognitive research and empirical AI*. New York, NY: Psychology Press.
Hoppe, E. I., & Kusterer, D. J. (2011). Behavioral biases and cognitive reflection. *Economics Letters, 110*(2), 97–100.
Jarmon, L., Traphagan, T., Mayrath, M., & Trivedi, A. (2009). Virtual world teaching, experiential learning, and assessment: An interdisciplinary communication course in Second Life. *Computers & Education, 53*(1), 169–182.
Jennett, C., Cox, A. L., Cairns, P., Dhoparee, S., Epps, A., Tijs, T., et al. (2008). Measuring and defining the experience of immersion in games. *International Journal of Human-Computer Studies, 66*(9), 641–661.
Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology, 3*(1), 1–24.
Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.
Larrick, R. P., Morgan, J. N., & Nisbett, R. E. (1990). Teaching the use of cost-benefit reasoning in everyday life. *Psychological Science, 1*(6), 362–370.
Lee, Y. H., Dunbar, N. E., Miller, C. H., Lane, B. L., Jensen, M. L., Bessarabova, E., et al. (2016). Training anchoring and representativeness bias mitigation through a digital game. *Simulation & Gaming*, 1–29.
Lewandowsky, S., Kalish, M., & Ngang, S. K. (2002). Simplified learning in complex situations: Knowledge partitioning in function learning. *Journal of Experimental Psychology: General, 131*(2), 163.
Lilienfeld, S. O., Ammirati, R., & Landfield, K. (2009). Giving debiasing away: Can psychological research on correcting cognitive errors promote human welfare? *Perspectives on Psychological Science, 4*(4), 390–398.
Martin, M. M., & Rubin, R. B. (1995). A new measure of cognitive flexibility. *Psychological Reports, 76*(2), 623–626.
McGonigal, J. (2011). *Reality is broken: Why games make us better and how they can change the world*. London: Penguin.
Milkman, K. L., Chugh, D., & Bazerman, M. H. (2009). How can decision making be improved? *Perspectives on Psychological Science, 4*(4), 379–383.
Morewedge, C. K., & Kahneman, D. (2010). Associative processes in intuitive judgment. *Trends in Cognitive Sciences, 14*(10), 435–440.
Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H., & Kassam, K. S. (2015). Debiasing decisions improved decision making with a single training intervention. *Policy Insights from the Behavioral and Brain Sciences, 2*(1), 129–140.
Mussweiler, T., Strack, F., & Pfeiffer, T. (2000). Overcoming the inevitable anchoring effect: Considering the opposite compensates for selective accessibility. *Personality and Social Psychology Bulletin, 26*(9), 1142–1150.
Nisbett, R. E., Fong, G. T., Lehman, D. R., & Cheng, P. W. (1987). Teaching reasoning. *Science, 238*(4827), 625–631.
Oechssler, J., Roider, A., & Schmitz, P. W. (2009). Cognitive abilities and behavioral biases. *Journal of Economic Behavior & Organization, 72*(1), 147–152.
Rassin, E. (2010). Blindness to alternative scenarios in evidence evaluation. *Journal of Investigative Psychology and Offender Profiling, 7*(2), 153–163.
Richey, M. K. (2013). *A game-based approach to teaching cognitive biases (Doctoral dissertation)*. Erie: Mercyhurst University.
Rowe, J. P., McQuiggan, S. W., & Lester, J. C. (2007). Narrative presence in intelligent learning environments. In *Paper presented at the 2007 AAAI fall symposium - technical Report, FS-07-05* (pp. 126–133). Arlington.
Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological Bulletin, 134*(2), 207.
Sinayev, A., & Peters, E. (2015). Cognitive reflection vs. calculation in decision making. *Frontiers in Psychology, 6*, 532.
Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*(1), 3.
Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology, 94*(4), 672.
Symborski, C., Barton, M., Quin, M., Morewedge, C. K., Kassam, K. S., & Korris, J. H. (2014). Missing: A serious game for the mitigation of cognitive biases. In *Paper presented at interservice/industry training, simulation and education conference, Orlando* (pp. 1–13).
Todd, B. (2014, October 7). *Sherlock Holmes: Crimes & Punishments review*. Retrieved March 23, 2017 from: https://www.gamespot.com/reviews/sherlock-holmes-crimes-punishments/1900-6415905/.
Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning, 20*(2), 147–168.
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131.
Van den Bosch, K., & Helsdingen, A. S. (2002). Improving tactical decision making through critical thinking. Proceedings of the Human Factors and Ergonomics Society. In *Paper presented at the human factors and ergonomics society 46th annual meeting, Baltimore, USA* (pp. 448–452). Thousand Oaks: SAGE Publications.
Veinott, E. S., Leonard, J., Lerner Papautsky, E., Perelman, B., Stankovic, A., Lorince, J., et al. (2013). The effect of camera perspective and session duration on training decision making in a serious video game. In *Paper presented at the games innovation conference international, Vancouver* (pp. 256–262). USA: IEEE Publications.
Wason, P. C. (1968). Reasoning about a rule. *The Quarterly Journal of Experimental Psychology, 20*(3), 273–281.
Yamnill, S., & McLean, G. N. (2001). Theories supporting transfer of training. *Human Resource Development Quarterly, 12*(2), 195–208.