Virtual Coach Application in Serious Games: Effects on Player Experience and Learning Outcomes

Master Thesis

Maarten Hoes – 6026907

Psychology Department, Utrecht University

Social, Health and Organisational Psychology

28th of July – Utrecht

Word Count: 8438

Supervisor: Dr Ruud Hortensius

Second Assessor: Dr Félice van Nunspeet

Abstract

In an effort to improve serious games, developers have increasingly started to implement virtual coaches. Many studies examining the impact of these coaches have found a positive effect on learning outcomes, but some studies failed to observe such an improvement. Research on whether virtual coaches beneficially affect the motivation and engagement of players is scarce. This research aims to contribute to the existing body of literature by assessing the effects of a virtual coach on both learning outcomes and player experience (i.e. motivation and engagement). The hypothesis is that the addition of a virtual coach improves these outcomes in a serious game aimed at teaching people how to effectively give feedback to others. A between-subjects research design was used. Participants were randomly allocated to one of two conditions: one group (n = 35) playing the serious game with a virtual coach and the other group (n = 35) playing without such a coach. Independent samples t-tests showed no significant effect of the coach on learning outcomes. In addition, no significant effect on motivation or engagement was observed. In accordance with previous studies, a significant negative correlation between age and learning outcomes was found, whereas in contrast to earlier research, females did not achieve better learning outcomes compared to males. This study further illustrates the well-known problem of the lack of generalisability regarding the findings of research examining virtual coaches. The limitations of this study concerning the application of the virtual coach are discussed extensively, culminating in recommendations of tangible steps that can be taken to improve the validity of similar research in the future.

Keywords: serious games, virtual coach, learning outcomes, motivation, engagement

Virtual Coach Application in Serious Games: Effects on Player Experience and Learning Outcomes

The current consensus is that technology-driven educational solutions such as e-learning and serious games are effective and of comparable quality to real life alternatives such as classes or seminars. This generally holds true for both online learning environments (Dankbaar, 2016; Dhawan 2020; Gitonga & Wambua, 2020; Reavley et al., 2018) as well as more specific technological learning solutions such as serious games (Girard et al., 2012; Zhonggen, 2019; Vogel et al., 2006; Schroeder et al., 2013). Application of virtual characters is an established practice within the field of (educational) games, and computer programmes in general (e.g. Schroeder et al., 2013). Think, for example, of the (in)famous paperclip that used to accompany Microsoft Word. The research on the effectiveness of e-learning and serious games has also delved into the application of virtual agents as coaches. According to the theory, players interact with this virtual coach as if it were human (Kim et al., 2007; Mayer et al., 2003). Through this mechanism, player motivation and engagement as well as ultimate learning outcomes would be affected positively by the presence of a virtual coach (Mayer et al., 2003). The positive effect of a virtual coach on learning outcomes has been confirmed in the majority of studies (Scholten et al., 2017; Schroeder et al., 2013). The effects of the application of a virtual coach on player engagement and motivation have been studied less often and findings so far have been contradictory (Scholten et al., 2017). This current research aims to contribute to the body of evidence by supplying proof of the effect of a virtual coach on both player experience (i.e. engagement and motivation) and learning outcomes. In the present study, an experiment will be designed and conducted to examine the effects on both learning outcomes and player experience of a virtual coach developed by the company of DialogueTrainer applied to their online learning platform aimed at improving conversational skills. To achieve this, literature on educational games and virtual coaches will first be examined. Consequently, the methods by which the experiment was conducted, and the data was analysed are established. Finally, the results and limitations of the present study will be interpreted and discussed.

Serious games can be defined as video games either based on (simulated) reality or fantasy whilst being entertaining, interactive, goal-oriented, and competitive with the primary aim of educating the user or player (Oprins et al., 2015). The latter part of this definition is where these games differ from normal video games, as there the primary goal is to entertain. An important note to make here is that serious games, as normal video games, can take vastly

different forms and structures. Accordingly, research findings in this field are often hard to generalise, as the experimental design of evaluative research often differs considerably across studies of different games (Zhonggen, 2019). This is a widely recognised problem in the academic field of serious games (e.g. Girard et al., 2012; Korteling et al., 2013; Oprins et al., 2015; Zhonggen, 2019). The most important question remains whether serious games can be realistic substitutes, alternatives, or additions to existing teaching and learning methods.

Results of research on the quality of educational serious gaming compared to traditional teaching methods vary considerably (Girard et al., 2012; Korteling et al., 2013; Scholten et al., 2017; Oprins et al., 2015; Zhonggen, 2019). This is partly due to the aforementioned problem of research findings being hard to generalise. Moreover, there is no consensus on the correct experimental procedures, as serious gaming and the body of research evaluating it is relatively new (Zhonggen, 2019). This is supported by the explosion of the number of publications on serious gaming since 2013, having increased by 150, which is effectively about 400% (Zhonggen, 2019). The difficulty with which results can be generalised effectively means that the effects and processes as postulated by theory have not yet been conclusively proven. However, meta-analyses of relevant studies reported generally positive effects on net learning outcomes (Girard et al., 2012; Vogel et al., 2006: Zhonggen, 2019).

Interestingly, meta-analyses examining the growing body of literature on serious games have begun to identify factors which significantly affect learning outcomes. These factors can roughly be split into two categories. Those related to characteristics of the player and those related to characteristics of the serious game. Player characteristics that have a significant effect of learning outcomes are age, gender, and unit size. Generally, research shows that younger players perform significantly better at serious game learning (Zhonggen, 2019), while female participants in generally outperform their male counterparts (Vogel et al., 2006; Zhonggen, 2019). Unit size here refers to the serious game being played in groups or individually, with individual playthroughs tending to me more effective (Vogel et al., 2006).

Many factors related to the serious game itself have been shown to influence learning outcomes. The realism of the serious game is commonly found to be of importance, as any unrealistic quality hampers player engagement with the learning material (Vogel et al., 2006; Zhonggen, 2019). It is noteworthy that realism here does not necessarily refer to photo-

realism or lifelike, but also to meeting user expectations. In the case of simulated conversations, this could take the form of having answer options corresponding to what the player would answer if the conversation were real. Another important factor is the backstory and debriefing of the player (Cheng et al., 2017). Backstory here takes the form of informing the player of the desired effects of the serious game before playing, where debriefing relates to supplying the players with feedback on their functioning (e.g. a score) (Zhonggen, 2019). When players are aware of the game's benefits and are given feedback, this improves the desired learning outcomes (Zhonggen, 2019). Also, serious games which are algorithmically adaptable generally achieve better learning outcomes (Schroeder et al., 2013; Vogel et al., 2006; Zhonggen, 2019). Thus, when the game allows for free input from the player and is able to realistically adapt and respond, more freedom for the player and a higher level of realism is achieved. Other factors positively affecting learning outcomes include clear instruction of how the game works and the ease of playing. (Iten & Petko, 2016; Tobias & Fletcher, 2007). Finally, the user interaction of the player with the game is of importance (Zhonggen, 2019). This refers to the player interface, but also to the use of animated agents when interacting with the player as well as the use of human voices over synthetic, computer-like ones (Tobias & Fletcher, 2007). One effective way in which several of the factors mentioned above can be incorporated in a serious game is by using virtual agents as coaches; coaches could feasibly increase realism by altering player expectations, help introduce the game, give feedback on performance, give instruction when needed, and could play an invaluable role in improving user interaction with the game.

A *virtual agent*, also referred to as a virtual character or pedagogical agent, is a digitally constructed character within a (serious) game which the player interacts with, or through (Schroeder et al., 2013). These characters, although generally human-like, can take a variety of forms. What is undeniably human-like about these virtual agents, however, is the way in which they interact with the player. These interactions can range from simple textual responses given by a static picture of a character to narrated words coming from a fully animated three-dimensional one, complete with non-verbal language. Generally, these virtual agents are not artificially intelligent, and therefore cannot adapt to custom player input (Scholten et al., 2017; Schroeder et al., 2013). The *virtual coach* essentially is a digital character that "helps participants make the most of an experience, in a way that breaks the fourth wall of any virtual situation" (Aldrich, 2009, p. 42). Their tasks can include introducing the game by providing background information, guiding the player through the

experience, giving feedback on their performance, and interpreting and presenting results (Aldrich, 2009).

The crux of the application of virtual agents is their interaction with the player, as the purpose is to engage and motivate the player (Aldrich, 2009; Girard et al., 2012; Oprins et al., 2015; Scholten et al., 2017; Schroeder et al., 2013; Veletsianos et al, 2009; Vogel et al., 2006). Poor interactions lead to feelings of anger, apprehension, and mistrust of the agent among the players which negatively affect their experience and learning outcomes (Veletsianos et al., 2009). With the surge of the use of virtual agents in pedagogical contexts, the body of literature examining human-agent interactions has also grown, showing that virtual characters are perceived as social agents by players (Scholten et al., 2017; Schroeder et al., 2013; Veletsianos et al., 2009). This means that the agents are treated as if they were human. According to this *Social Agency* theory, when confronted by a virtual agent, players implicitly recognise it as being a social situation, and apply the same rules as during human-human interactions (Schroeder et al., 2013; Kim et al., 2007). Seeing the digital character as a social partner improves selecting, organising, and integrating information as these processes also happen in human-human interactions (Mayer et al., 2003). This is supported by research showing that players prefer agents who display flattering texts compared to agents who do not, and agents who are arbitrarily included in the player's team (Nass & Brave, 2005). These findings are all concurrent with how humans judge other humans in similar situations.

Research examining the effects of virtual agents in serious games has so far mainly focused on learning outcomes while their effect on player experience has rarely been explored (Scholten et al., 2017). Generally, research has found that programmes applying virtual agents have a small but significant positive effect on learning outcomes compared to programmes without such agents (Schroeder et al., 2013). Several characteristics of virtual agents have been found to positively affect learning outcomes. The use of human voices compared to no-voice or computer-generated voice alternatives improve learning outcomes (Mayer et al., 2003). Also, animated agents able to express emotions in a complex manner, positively affect learning outcomes (Schroeder et al., 2013), even if only the face is animated (Sproull et al., 1996). The positive effect of animation supports the Social Agency theory. Further proof for this comes from research showing that players effectively build a relationship with coaching agents during interactions as measured by levels of perceived working alliance and rapport between players and virtual agents (Gratch et al. 2007; Bickmore & Picard, 2005). The reported levels were comparable to face-to-face (human)

interactions. The tendency of players to regard agents as fellow humans is also evident from the fact that virtual coaches that are responsive are superior to unresponsive coaches (Scholten et al., 2017). Here, responsive and unresponsive refers to the capacity of agents to react appropriately to input by the player. This input can entail a wide variety of signals, ranging from simple text to visual and audio cues, to which the responsive virtual agent then responds in an emotionally appropriate way. Veletsianos et al. (2009) highlight the importance of focusing on user interaction, the message, and the characteristics of the virtual coach. Respectively, these foci relate to agents being attentive to and gratifying the player's needs, agents clearly communicating whilst taking into account the player's experience, and agents displaying socially appropriate behaviour. Most characteristics discussed in this paragraph echo the serious game-related characteristics shown to positively affect learning outcomes.

Few studies examined the effect of virtual coaches on player motivation and engagement. One study examining the effect of a 'gaze intelligent tutoring system' found that effects on motivation and engagement were minimal (D'Mello et al., 2012). This gaze intelligent tutoring system referred to a virtual agent capable of reacting to perceived inattentiveness on the part of player by using eye tracking technology and attempted to reengage the player when they looked away from the screen (D'Mello et al., 2012). Another study supplied proof that the use of an animated virtual coach improved engagement compared to non-animated substitutes such as arrows pointing to relevant information (Van Mulken et al., 1998). The best part of the available literature on player experience, however, focused on the usage of animated agents in general, rather than on agents fulfilling the role of virtual coach (Dehn & Van Mulken, 2000).

To contribute to filling this gap in the literature, in the current research the effect of a virtual coach on motivation and engagement will be measured, in addition to the effect on learning outcomes. In view of the reviewed literature, a well-crafted virtual coach guiding players through a serious game should increase the sense of realism by being more human in its behaviour, introducing the game effectively by clearly communicating, providing custom feedback based on player input, and generally improving user interaction with the application by forming a relationship with the player. This should increase user engagement and motivation as well as improve learning outcomes. Based on this, it was hypothesised that player experience and learning outcomes are positively affected by the addition of a virtual coach, exhibiting some of these characteristics, in a serious game designed to improve

conversational skills. In this study, player experience consists of measures of motivation and engagement with the game. Learning outcomes are operationalised as the score achieved by players on the internal scoring system of the serious game.

Methods

Participants

Participants had to be fluent in Dutch, and able to work with computers by themselves without requiring additional aid. Participants were recruited through a variety of channels including Facebook, WhatsApp, and LinkedIn. Additionally, snowball sampling was used. In order to optimize recruitment, participants were offered free access for three months to a similar game on the DialogueTrainer platform (the platform used in the experiment), after the study was completed. No formal sample size calculations were made, since the effects of the virtual coach on the outcome parameters are unknown. Earlier studies applying somewhat similar interventions and comparable outcome measurements typically included 50-70 participants (Schroeder et al., 2013). Based on this, a sample size of 70, or 35 per condition was aimed for. Before starting data collection, ethics approval for this experiment was obtained from the Utrecht University Faculty Ethics Review Board (FERB) and participants were required to give their informed consent before starting the experiment.

Experimental design

In order to determine the effects that a virtual coach has on the user experience and performance a between-subjects research design was used. Participants were randomly assigned to one of two comparison groups using the flow function of the Qualtrics Survey website. One group played a serious game with the virtual coach and the other group without virtual coach. Firstly, players were introduced to the aim of the game, which was to learn how to give feedback correctly. After this, players received the theory they needed to do this correctly before playing a simulation exercise where they applied what they learned. As part of this simulation, players received direct feedback on their performance. The final part of the module consisted of a short reflection exercise. For the group with the virtual coach, the introduction to the goal and use of giving feedback (i.e. contextual information) and the reflection exercise were given by a virtual coach. In the control group without the virtual coach, these were achieved by simple slides of text. To assess the outcome parameters of this study, all participants, after having finished the module, were re-directed to the final part of the Qualtrics survey, the link to which was integrated into the DialogueTrainer platform.

Materials and outcome parameters

The serious game which was used for this experiment is the simulation-based game developed by the company of DialogueTrainer. This company offers general and tailor-made simulated conversations to institutions looking for ways to train their professionals or students in a variety of communicational skills. Generally, within these simulated conversations, first a case and computer-avatar are introduced to the player, with the corresponding theory and goals to achieve (e.g. delivering bad news quickly rather than waiting). Then, the conversation starts and the simulation takes the form a branching path structure, with options chosen by the player resulting in corresponding reactions of the computer which the player then reacts to again (see Appendix A for example). After the simulated conversation comes to an end, feedback is given to the player on a variety of parameters depending on the simulation (e.g. directness in a bad-news simulation). The simulation exercise referred to within this section refers to such a simulated conversation. The company also offers more comprehensive experiences in the forms of *modules*, where an introduction, theory, and reflection are integrated with the simulation exercise. Such a module with the topic of giving feedback was used as the serious game in this study (see Appendix A for more information on the module used). Recently, in order to improve the player experience and the effects of their products, DialogueTrainer has developed virtual characters to act as coaches during some simulations.

The virtual coach used within this study has been created specifically to suit the 'giving feedback' module. According to the developers, the DialogueTrainer virtual coaches are created with the aims of "introducing the subject of the simulation, giving feedback, building a relationship with the player, and allowing for space to emotionally interact with the player" (M. Hulsbergen, personal communication, June 9, 2021). Within the context of this experiment, the script of the virtual coach was kept as close to the original text of the module as possible. This refers to the two parts that were conducted by the virtual coach, namely the introduction (giving contextual information) and the reflection. However, some small alterations between the texts were made so as to make the interaction with the virtual coach appear more natural. The control group was only offered these parts in the form of the original texts (see Appendix B for more information on the functions of the virtual coach and the differences between the experimental groups).

Learning outcomes were measured through the internal scoring system of DialogueTrainer simulations. These scoring systems are created individually for each simulation and measure the performance of students in a particular exercise on a scale, ranging from 0 to 100. In practice, a 100 score means that the optimal answers, according to the theory underlying the simulation, were chosen by the player. Lower scores indicate that players made sub-optimal choices and are indicative of an incomplete understanding of the same theory. Participants were asked to only play the simulation exercise once. Any scores gained by participants in consecutive playthroughs were ignored. This method was chosen as it was presumed that alternatives (e.g. calculating average scores over a larger number of playthroughs) would result in biased results favouring those respondents who had more time to participate in the experiment and were thus able to play the simulation until they perfected it.

Player motivation was measured through using the Intrinsic Motivation Inventory (IMI), translated to Dutch (Vernieuwenderwijs, 2021). The IMI was selected as it often used to measure the motivation that people feel towards a certain activity. The scale consists of 22 items for which the participants provide a rating on the extent that they agreed with a statement concerning how motivated, competent, and pressured they felt during the activity (e.g. “I did not feel nervous during the activity”). Responses are rated on a scale from 0 (*not at all true*) to 7 (*very true*). The reliability of the IMI scale within this study was acceptable, Cronbach’s α = .79, though somewhat lower when compared with previous findings (Ostrow & Heffernan, 2018).

Player engagement was measured using the Short Form-User Engagement (SF-UE) which is often used in human-computer interactions (O’Brien et al., 2018). No localised version of this scale was available in Dutch, so it was translated by a native-level speaker. The scale consists of 12 items for which the participants provide a rating on to what extent they agreed with a statement concerning if they were engaged by the activity (e.g. “the time I spent playing the simulation, flew past”). Responses are rated on a scale from 0 (*strongly disagree*) to 5 (*strongly agree*). The reliability of the SF-UE within this study was good, Cronbach’s α = .80, which is comparable to previous findings (O’Brien et al., 2018). For the complete questionnaires used in this study see Appendix C.

Data Analysis

The data was compiled and analysed using SPSS Statistics version 26. First, the username of participants was used to link the score achieved on the DialogueTrainer platform to the answers of the questionnaire. After this, the data was anonymised. In order to determine the effect of the virtual coach, independent samples t-tests were conducted for the dependent variables learning outcomes, motivation, and engagement. In addition, Pearson product moment correlation coefficients were calculated for all model variable combinations, except for the categorical variable gender, where point-biserial correlation coefficients were determined. Finally, an analysis of covariance (ANCOVA) was applied to assess whether any effect of the virtual coach on the outcomes could be explained by differences in age between the two groups, since age is known to have a significant influence on learning outcomes and the data suggested such an effect might be present in the sample.

Results

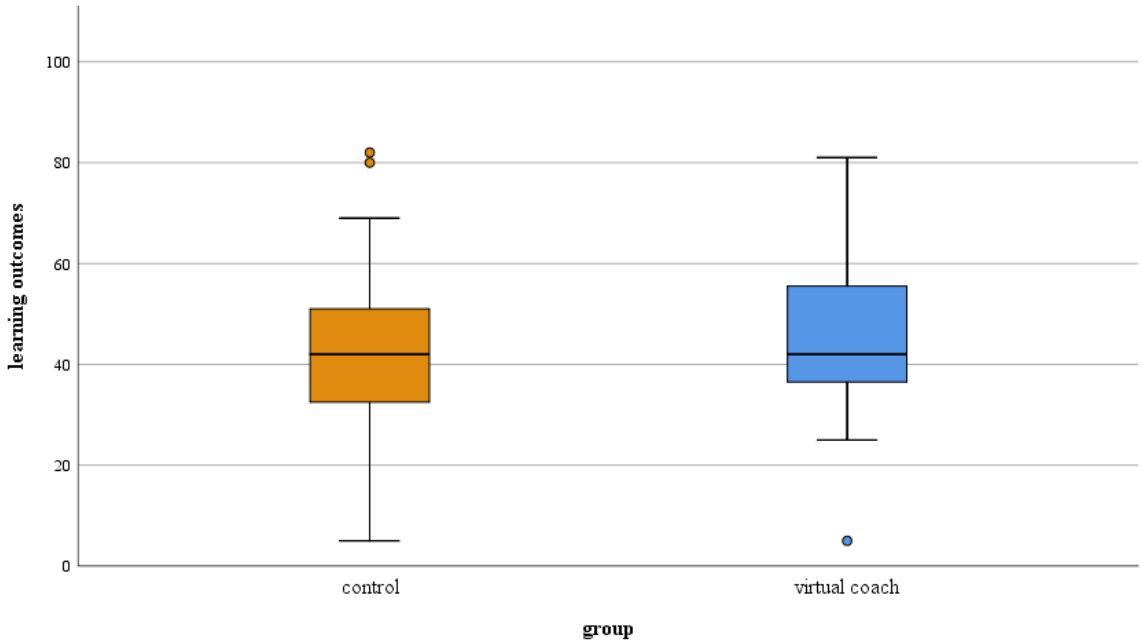
A total of 78 people participated in this study, of which 70 completed the entire process. Eight participants started the study and entered the DialogueTrainer platform but did not finish the experiment. Seven of these did not play any simulations at all whilst one participant did play the simulations but did not complete the final questionnaire. This group was comprised of four people of each condition with the one having played the simulations belonging to the virtual coach group. In the final sample, the virtual coach group consisted of 35 people and the control group included 35 people. The overall age of all participants ranged from 18 to 70 years (M = 35.20, SD = 16.39). In the virtual coach group the age ranged from 20 to 65 years (M = 34.37, SD = 16.52), and in the control group from 18 to 70 years (M = 35.83, SD = 16.47). In the final sample 35 participants identified themselves as being male and 35 as being female, with the virtual coach group having 17 male and 18 female participants and the control group having 18 male and 17 female participants.

In order to address the main research question, the data on the three dependent variables was first explored. Shapiro-Wilk tests of normality showed normal distributions in the virtual coach group for learning outcomes, $W(35) = 0.98$, $p = .646$, for motivation, $W(35) = 0.98$, $p = .626$, and for engagement, $W(35) = 0.97$, $p = .486$. In the control group data was also normally distributed for learning outcomes, $W(35) = 0.96$, $p = 206$, for motivation, $W(35) = 0.97$, $p = .510$, and for engagement, $W(35) = 0.94$, $p = .068$. In Figures 1 to 3, the data of both groups are visualised for learning outcomes, motivation, and engagement,

respectively. The potential outliers depicted in these figures were not excluded from the data analysis as the assumption of normality was not violated. In Figure 1 it can be seen that learning outcomes were generally not very high whilst the total range in scores was large. An independent samples t-test showed that there was no statistically significant difference in learning outcomes between the virtual coach (M = 44.91, SD = 14.72) and control (M = 42.34, SD = 17.13) conditions, $t(68) = -0.67$, $p = .503$, although the virtual coach group did score somewhat higher than the control group. Figures 2 and 3 show that people were generally motivated and engaged by the module in both conditions as both means were higher than the middle point of the Likert scales, being 4 and 3 respectively. For player motivation, no significant difference was found between the virtual coach (M = 4.29, SD = 0.66) and control (M = 4.50, SD = 0.62) conditions, $t(68) = 1.37$, $p = .177$. Finally, the last independent samples t-test showed there was no significant difference in player engagement between the virtual coach (M = 3.59, SD = 0.49) and control (M = 3.69, SD = 0.56) conditions, $t(68) = 0.83$, $p = .407$. For both player experience outcomes the virtual coach group reported slightly lower values than the control condition.

Figure 1

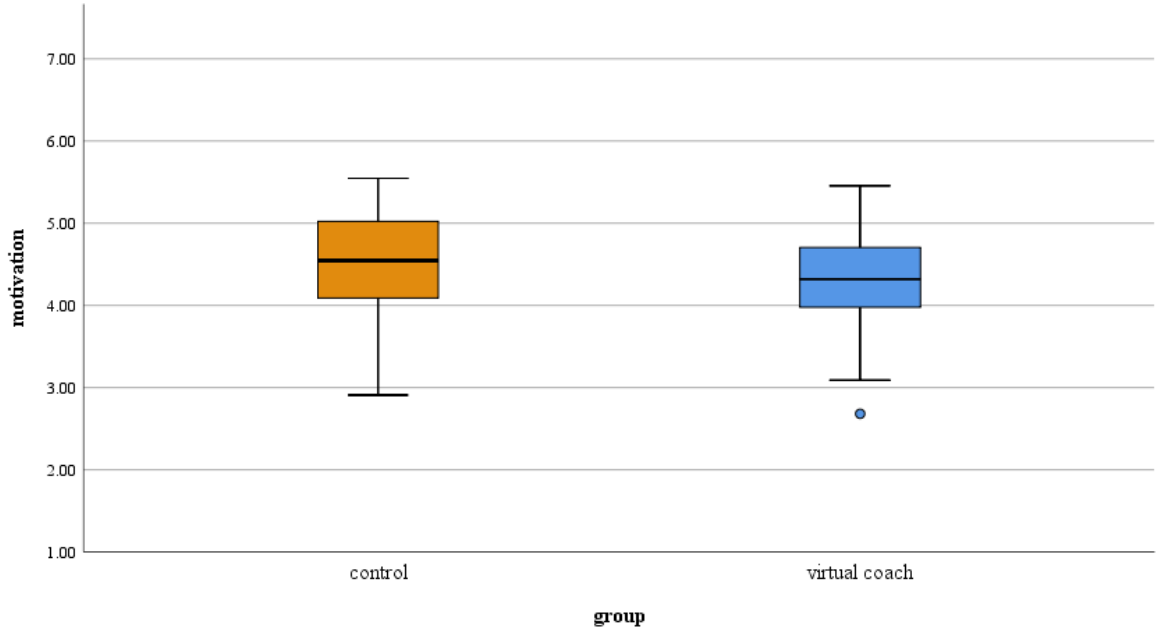
Boxplot Showing the Distribution of Learning Outcomes for Both Groups



Note. n = 35 for both groups. Individual points denote potential outliers within the data set.

Figure 2

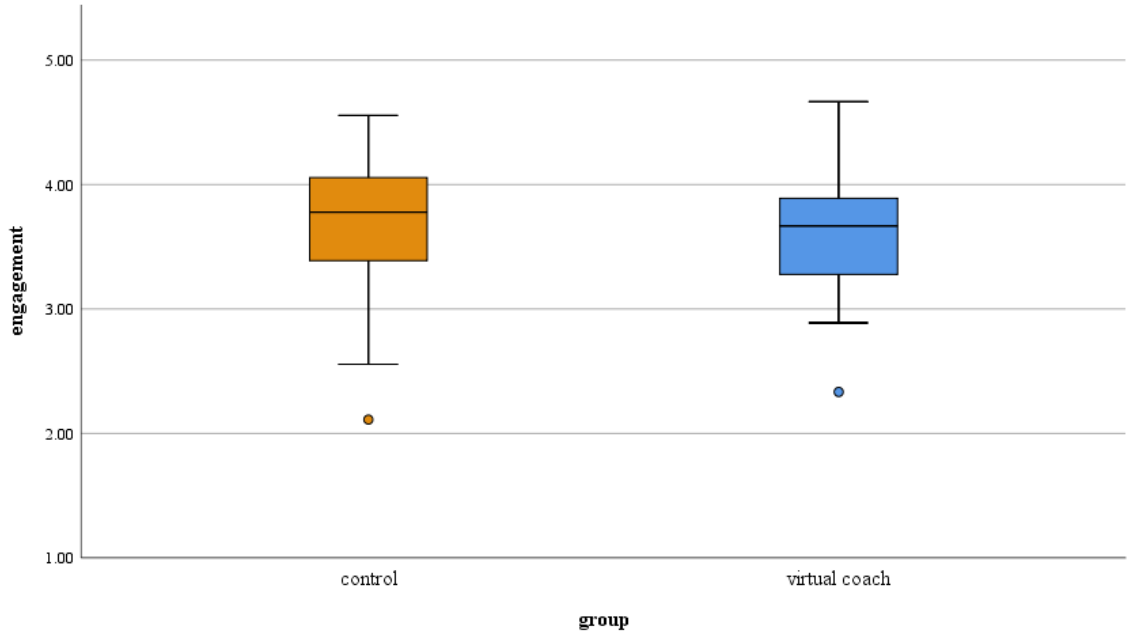
Boxplot Showing the Distribution of Motivation for Both Groups



Note. n = 35 for both groups. Individual points denote potential outliers within the data set.

Figure 3

Boxplot Showing the Distribution of Engagement for Both Groups



Note. n = 35 for both groups. Individual points denote potential outliers within the data set.

As evident from the literature research, previous studies have shown that player characteristics such as gender and age can have a significant effect on learning outcomes. To determine whether these particular effects were present in the sample data, as well as whether other interactions between model variables were significant, Pearson product-moment correlation coefficient were computed between all model variables except gender, for which point-biserial correlation coefficients were computed with all other variables (Table 1). A significant correlation was found between age and learning outcomes, $r = -0.24$, $p = .048$, but not between gender and learning outcomes, $r = -0.65$, $p = 592$. Additionally, a significant correlation was found between motivation and engagement, $r = 0.63$, $p < .000$. Figure 1 shows the relation found between age and learning outcomes. The small negative effect means that, across the whole sample, participants that were older tended to score slightly worse. The relatively strong correlation between motivation and engagement shows that participants who felt more motivated were generally also more engaged with the simulation.

Table 1

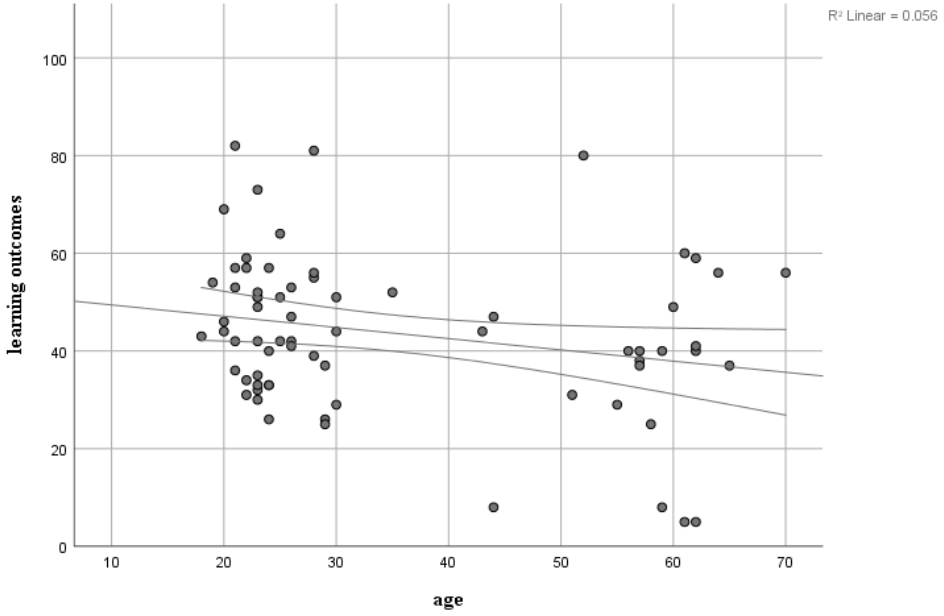
Computed Correlations for Model Variables

Variable	1	2	3	4	5
1. Age	-				
2. Gender	0.01	-			
3. Learning outcomes	0.24*	-0.07	-		
4. Motivation	1.03	-0.74	0.13	-	
5. Engagement	-0.02	-0.02	0.12	0.63**	-

Note. * $p < .05$ ** $p < .01$, $N = 70$.

Figure 4

Scatterplot Showing Correlation Between Age and Learning Outcomes



Note. Upper and lower lines show 95% CI. $N = 70$.

Following this finding, an ANCOVA was conducted to determine whether the observed insignificant effect of the virtual coach on learning outcomes found in the independent samples t-test could be attributed to the effect of age on learning outcomes, albeit that as both groups were randomly assigned and Levene's test of variances showed that age was evenly distributed among both conditions, $F(1,68) = 2.12$, $p = .647$, it is unlikely that age would be a significant moderator. Indeed, after controlling for age, still no significant difference was found between the learning outcomes of the virtual coach group ($M = 44.77$, 95% CI [39.49, 50.05]) and the control group ($M = 42.49$, 95% CI [37.21, 47.76]), $F(1, 67) = 0.37$, $p = .543$.

Discussion

The main aim of this study was to determine whether the application of a virtual coach to the DialogueTrainer platform would positively affect learning outcomes and player experience. On average, the virtual coach group did score slightly better on learning outcomes than the control group, but not significantly so. The player experience of the virtual coach group was generally somewhat worse as players felt both less engaged and less motivated towards the module than the control group, although this difference was again far from statistically significant. A significant negative correlation was found between age and

learning outcomes, but not between age and gender. Having established the effect of age on learning outcomes within the sample, after factoring this into the effect of the virtual coach on learning outcomes a significant difference between conditions was still undetectable. Finally, a significant correlation was found between the dependent variables of player motivation and engagement.

Eight out of 78 participants did not complete the experiment. Further analysis of when these participants quit the study showed no differences between the two groups. Thus, no relevant conclusions can be drawn about the effect of the virtual coach on experiment completion. Roughly ten percent of participants decided to leave the experiment upon entering the DialogueTrainer platform after having been thoroughly explained what to expect in the first part of the experiment. This suggests that some people find the platform itself, or interaction with it, daunting or complicated. No firm conclusions can be made about this, however, nor about the characteristics of these participants, as no data was collected at the point of leaving.

The lack of a clear effect of the virtual coach on learning outcomes observed in our study is in contrast with findings of most other studies, where the presence of a virtual coach had a positive effect (Scholten et al., 2017; Schroeder et al., 2013). The slightly higher scores on learning outcomes in the virtual coach group were far from statistically significant and such small differences are unlikely to be relevant. Importantly, it is not uncommon either for applications of virtual coaches in serious games to confer no detectable effects (Girard et al., 2012; Korteling et al., 2013; Oprins et al., 2015; Zhonggen, 2019). This is often attributed to the large variance in virtual coaches and their application. The virtual coach developed by DialogueTrainer which was used in this study is unique, since the code was written by the company itself and is not used in this particular way by anyone else. Consequently, findings of research applying different coaches might not be applicable to the coach used in the context of this study. Several virtual coach characteristics that significantly affect learning outcomes have still been identified in the literature. Whether or not these are present in the DialogueTrainer virtual coach could be instrumental in explaining the absence of a significant effect. The use of a human voice during interaction (Mayer et al., 2003) and the use of facial expression to convey emotional information (Schroeder et al., 2013; Sproull et al., 1996) were both exhibited by the virtual coach in this study. The presence or absence of other important characteristics reported to affect virtual coach outcomes, such as being attentive to and gratifying the player's needs, agents clearly communicating while taking into account the

player's experience, and agents displaying socially appropriate behaviour (Veletsianos et al., 2009) was harder to determine, as these are more subjective and dependent on the creation process of the virtual coach itself. Considering both the design philosophy of DialogueTrainer behind these coaches as well as the dialogue options between the player and the coach it seems that the coach does exhibit these characteristics to a reasonable degree (see Appendix B). An attribute that is missing in the virtual coach assessed in this experiment, however, is that of the responsive agent. If a virtual coach is able to respond to custom player input, this can result in better learning outcomes (Scholten et al., 2017). The manner in which DialogueTrainer simulations are made (via pre-written dialogue options) does not allow for such responsiveness. In order to include this attribute, some form of artificial intelligence would have to be used which is often very expensive and complicated (Scholten et al., 2017). Thus, most of the effective characteristics of virtual coaches established by previous research were present in the experiment, implying that the attributes of the coach are unlikely to explain the lack of a detectable effect on learning outcomes.

The lack of an observed effect could theoretically also be attributable to differences in age and gender between the two groups as these variables are known to influence learning outcomes. It is generally found that players that are female and younger have better learning outcomes when playing serious games (Zhonggen, 2019). However, these relations were not ubiquitously present in the sample data. A significant negative correlation was found between age and learning outcomes, echoing previous findings that younger people generally outperform older people in serious games, but the finding that female players outperform male players was not supported by the present data. After having established the correlation between age on learning outcomes within the sample, a significant difference between conditions was still undetectable when factoring age into the effect of the virtual coach on learning outcomes. This was expected, as participants were randomly allocated to the two conditions and, thus, no differences in age were observed between the two groups. There is no definite theory which explains either the relation between learning outcomes and age or that between learning outcomes and sex (Zhonggen, 2019). The respective presence and absence of these relations in this study can therefore not be adequately explained. However, it can be posited that older people find it relatively harder to interact effectively with digital programmes such as the DialogueTrainer platform. The lack of a clear effect of the virtual coach on learning outcomes could therefore be partly explained by the fact that the current sample included a relatively large proportion of older individuals, as is visible in Figure 4.

According to the literature, the virtual coach should positively affect learning outcomes and player experience primarily through its interaction with the player (Aldrich, 2009; Girard et al., 2012; Oprins et al., 2015; Scholten et al., 2017; Schroeder et al., 2013; Veletsianos et al, 2009; Vogel et al., 2006). As has already been discussed, the lack of an effect is likely not attributable to the various characteristics of the virtual coach itself. Therefore, the manner in which the coach was used within the context of this experiment must be explored further. The Social Agency theory posits that the virtual coach is considered to be human by the player, and is interacted with accordingly (Kim et al., 2007). During interaction with social partners, the occurrence of selecting, organising, and integrating information increases (Mayer et al, 2003).

Applying this to the context of the current study, it is implied that any information given to the player by the virtual coach can generally be remembered better by participants than when such information is given simply via text. Participants in the virtual coach condition are then able to better use this information during the simulation exercise, and thus achieve a higher score. The only components of the module which were different between the two conditions were the introduction and reflection parts (see Appendix B). Out of these, the introduction was the only which provided contextual information that could have aided the player in achieving a higher score. Notably, within the simulated interaction with the virtual coach players were given the option to forego additional information and proceed directly to the theory. In the control group, this information was given in an introductory text which could also be skipped. It could be that a large proportion of participants in the virtual coach group skipped this information. As a result, they would have interacted with the virtual coach for a shorter period, as well as received less information. Both of these factors can lead to a decline in scores in the simulation exercise, and this could thus, at least partly, explain the lack of a discernible effect of the virtual coach on learning outcomes. The relatively shorter amount of time that participants spent with the virtual coach could also explain why participants in the virtual coach condition were not more motivated or engaged. Consequently, part of the absence of a clear effect of the virtual coach on all outcome variables could be explained by limited contrast between the two comparison groups in the amount of information received or in the way in which this information was delivered. Data to confirm which information the virtual coach group received was not available, however, nor was it possible to determine whether or not the control group skipped or read the additional information.

Finally, the absence of a significant difference in mean learning outcomes between the groups could be attributed to the operationalisation of learning outcomes within this study. Learning outcomes are often best measured longitudinally via interviews, but this was not possible within the limited time frame of this study (Wiggins, 1998). Moreover, an eligible substitute was already present in the form of the internal scoring system of the DialogueTrainer platform. In this system, players can receive a score ranging from 0 to 100 reflecting the quality of their choices during a simulation exercise. Having a better understanding of the underlying theory and context will lead to better scores. The choice for applying the internal scoring system of the DialogueTrainer platform to operationalise learning outcomes is justifiable in this sense, as it does reasonably reflect the grasp of understanding of the player.

There are two factors, however, which could decrease the validity of this outcome measurement. Firstly, the scoring system is tailored individually by each DialogueTrainer developer to each simulation exercise. An interview with the developers on how these scorings systems are formed resulted in various explanations. As a result of this somewhat unclear formation of scoring systems, it cannot always be assumed that a better understanding of the theory (or learning outcomes) necessarily leads to a higher score. However, the simulation used in this experiment has been used by many DialogueTrainer players before and has been improved multiple times using the players' feedback. This particular scoring system can therefore be expected to be generally good in this regard. Secondly, the operationalisation does not take into account the manner in which players prefer to learn. Specifically, participants were asked to only play the simulation once. This is generally not the manner in which people use the DialogueTrainer platform. Rather, they can exhibit a variety of different learning profiles, and often tend to play the simulation multiple times, resulting in varying scores (Mendes, 2021). However, alternative measurements of learning outcomes such as calculating an average over multiple playthroughs were deemed to be unfairly advantageous to those participants who had more time to spare, while simultaneously making it nearly impossible to provide participants with an accurate estimate of the time it would take to complete the experiment. For these reasons, only the first achieved score was used in the data analysis, even though many participants nevertheless ended up playing the simulation multiple times. Additionally, the simulations are constructed in such a way that early failure of the exercise is possible, ending the scenario early and leaving the player with a low score. In practice, this can mean that a relatively small lap in judgement can result in a

score as low as 5, while this does not necessarily accurately reflect the learning outcomes for the player. Data showing that scores under 10 did occur seem to support this idea.

In order to overcome the problems regarding the learning outcomes discussed above, it is recommended that future research on the topic of virtual coaches, or more specifically on the platform of DialogueTrainer, should take various factors into account. Firstly, researchers should preferably aim to create their own virtual coach. Rather than using a virtual coach created by a company to suit their platform or modelling the coach after an existing one, this coach should be based on targeted research taking the specific topic and aim of the anticipated study into account. This can help to overcome the generalisability problem of research findings on virtual coaches. More specifically, the EnALI framework can be used, which was created specifically for this purpose (Veletsianos et al., 2009). Secondly, care should be taken to create enough contrast between the experimental virtual coach and the control group. In the present study it is conceivable that the relatively small role the virtual coach played, as it only replaced two components compared to the control group, at least partly explains our findings. These components equalled to about a third of the total time the experiment took. The influence of the virtual coach on the player experience might therefore be smaller than it could have been, had the coach been assigned a more prominent role. This can be achieved in future research by, for example, explicitly phasing the virtual coach in and out during the experiment. In the context of this study, this could have been achieved by having the virtual coach give feedback on the player performance during the simulation exercise. The options in this regard are naturally limited by the technological capabilities of the platform used. In addition, it is recommended that future research considers to include more operationalisations of learning outcomes than chosen in the present study, ideally by interviewing participants longitudinally. Finally, the use of scores calculated over multiple playthroughs could provide additional information, but whether this approach truly reflects the effect of virtual coaches in daily practice will have to be researched further.

Possible reasons behind the lack of a significant effect on learning outcomes within this study discussed above, could also play a role in the interpretation of the results regarding the effect of the virtual coach on player experience. As earlier studies established that a virtual coach generally positively affects learning outcomes (Scholten et al., 2017; Schroeder et al., 2013), the lack of a significant effect on player experience can therefore not be taken as proof that a virtual coach does not affect player motivation and engagement. Determining the existence of such an effect was one of the aims of this research, and due to the same

limitations described above, definite conclusions cannot be drawn. Future research aiming to explore this topic further, should take into account the recommendations outlined above, aimed at improving the rigidity and validity of experiments using virtual coaches.

What can be concluded regarding player experience, is that a strong correlation was found between the two dependent variables engagement and motivation, which comprised the construct of player experience within the context of this experiment. This observed correlation is not unexpected, as it is logical that those people who feel more engaged by an activity also feel more motivated towards it. Nonetheless, it does lend merit to the bundling of these two constructs as player experience in this research. More specifically, this correlation showing that engaged people were generally also motivated and vice versa can be considered tentative proof that these two constructs are affected by serious games in the same manner. Serious games positively affecting both motivation and engagement has been found in previous research (Aldrich, 2009; Girard et al., 2012; Oprins et al., 2015; Scholten et al., 2017; Schroeder et al., 2013; Veletsianos et al, 2009; Vogel et al., 2006). The present study's finding shows that this relation is not nullified by the addition of a virtual coach.

In conclusion, the present study was unable to confirm the hypothesis that learning outcomes of players would be positively affected by the addition of a virtual coach on the DialogueTrainer platform. This is in contrast with previous research showing that learning outcomes are improved by using such agents. Also, player experience was not clearly affected by adding a virtual coach. Due to several limitations of the current study, no definitive conclusions regarding the effect of a virtual coach on learning outcomes and player experience can be drawn. These limitations were discussed at length, as well as recommendations on how to overcome these. By applying these to future research, it can become clearer if and in what way virtual coaches positively affect both the learning outcomes and player experience of digital educational solutions such as serious games.

References

Aldrich, C. (2009). *Learning online with games, simulations, and virtual worlds: Strategies for online instruction* (Vol. 23). John Wiley & Sons.

Bickmore, T.W., & Picard, R.W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction*, *12*(2), 293–327. https://doi.org/10.1145/1067860.1067867

Cheng, M.-T., Lin, Y.-W., She, H.-C, & P.-C. Kuo. (2017). Is immersion of any value? Whether, and to what extent, game immersion experience during serious gaming affects science learning. *British Journal of Educational Technology*, *48*(2), 246-263. https://doi.org/10.1111/bjet.12386

Dankbaar, M. (2016). Serious games and blended learning; effects on performance and motivation in medical education. *Perspectives on Medical Education*. *6*. https://doi.org/10.1007/s40037-016-0320-2

Dehn, D. M., & Van Mulken, S. (2000). The impact of animated interface agents: a review of empirical research. *International journal of human-computer studies*, *52*(1), 1-22. https://doi.org/10.1006/ijhc.1999.0325

D'Mello, S., Olney, A., Williams, C., & Hays, P. (2012). Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of human-computer studies*, *70*(5), 377-398. https://doi.org/10.1016/j.ijhcs.2012.01.004

Dhawan S. (2020). Online Learning: A Panacea in the Time of COVID-19 Crisis. *Journal of Educational Technology Systems*, *49*(1), 5-22. https://doi.org/10.1177/0047239520934018

Girard, C., Ecalle, J., & Magnan, A. (2012). Serious games as new educational tools: how effective are they? A meta-analysis of recent studies. *Journal of computer assisted learning*, *29*(3), 207-219. https://doi.org/10.1111/j.1365-2729.2012.00489.x

Gitonga, M. O., & Wambua, A. W. (2020), Effectiveness of eLearning During the Lockdown Period-Kenya Case Study. In Hamasi, L., & Amutabi, M. (Eds.), *Politics and Sustainable Development in Africa*. (pp. 30-47). https://www.academia.edu/download/64029090/covid-pandemic-scopus.pdf

Gratch, J., Wang, N., Gerten, J., Fast, E., Duffy, R. (2007). Creating rapport with virtual agents. In: Pelachaud, C., Martin, J.C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) *IVA 2007. (LNAI, 4722,* 125–138. Springer, Heidelberg. https://doi.org/10.1007/978-3-540-74997-4_12

Iten, I., & Petko, D. (2016). Learning with serious games: Is fun playing the game a predictor of learning success? *British Journal of Educational Technology*, *47*(1), 151-163. https://doi.org/10.1111/bjet.12226

Kim, Y., Baylor, A. L., & Shen, E. (2007). Pedagogical agents as learning companions: The impact of agent emotion and gender. *Journal of Computer Assisted Learning*, *23*(3), 220-234. https://doi.org/10.1111/j.1365-2729.2006.00210.x

Korteling, J. E., Helsdingen, A. S., Theunissen, N. C. M. (2013). Serious Games @ Work: Learning job-related competences using serious gaming. In A. Bakker & D. Derks (Eds.), *The Psychology of Digital Media at Work* (pp. 123-144). London, UK: Psychology Press LTD / Taylor & Francis Group. https://www.taylorfrancis.com/chapters/edit/10.4324/9780203074145-12/serious-gaming-work-learning-job-related-competencies-using-serious-gaming-hans-korteling-anne-helsdingen-nicolet-theunissen

Mayer, R. E., Sabko, K., & Mautone, P. (2003). Social cues in multimedia learning: Role of speaker's voice. *Journal of Educational Psychology*, *95*(2), 419-425. https://doi.org/10.1037/0022-0663.95.2.419

Mendes, A. (2021). *Learning strategies in an online educational simulation* [Master's thesis, Utrecht University].

Nass, C. I., & Brave, S. (2005). *Wired for speech: How voice activates and advances the human-computer relationship* (1st ed.). Cambridge, MA: MIT press.

O'Brien, H. L., Cairns, P., & Hall, M. (2018). A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies*, *112*, 28-39. https://doi.org/10.1016/j.ijhcs.2018.01.004

Oprins, E., Visschedijk, G., Roozeboom, M. B., Dankbaar, M., Trooster, W., & Schuit, S. C. (2015). The game-based learning evaluation model (GEM): measuring the

effectiveness of serious games using a standardised method. *International journal of technology enhanced learning*, *7*(4), 326-345. https://doi.org/10.1504/IJTEL.2015.074189

Ostrow, K. S., & Heffernan, N. T. (2018). Testing the validity and reliability of intrinsic motivation inventory subscales within assistments. *International conference on artificial intelligence in education*, *5*(2), 381-394. https://doi.org/10.1007/978-3-319-93843-1_28

Sproull, L., Subramani, M., Kiesler, S., Walker, & J.H., Waters, K. (1996). When the interface is a face. *Human Computer Interaction*, *11*(2), 97–124. https://doi.org/10.1207/s15327051hci1102_1

Reavley, N. J., Morgan, A. J., Fischer, J. A., Kitchener, B., Bovopoulos, N., & Jorm, A. F. (2018). Effectiveness of eLearning and blended modes of delivery of Mental Health First Aid training in the workplace: randomised controlled trial. *BMC psychiatry*, *18*(1), 1-14. https://doi.org/10.1186/s12888-018-1888-3

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology*, *25*(1), 54-67. https://doi.org/10.1006/ceps.1999.1020

Schroeder, N. L., Adesope, O. O., & Gilbert, R. B. (2013). How effective are pedagogical agents for learning? A meta-analytic review. *Journal of Educational Computing Research*, *49*(1), 1-39. https://doi.org/10.2190/EC.49.1.a

Scholten, M. R., Kelders, S. M., & van Gemert-Pijnen, J. E. (2017). A scoped review of the potential for supportive virtual coaches as adjuncts to self-guided web-based interventions. *International Conference on Persuasive Technology* (pp. 43-54). Springer, Cham. https://doi.org/10.1007/978-3-319-55134-0_4

Tobias, S., & Fletcher, J.D. (2007). What research has to say about designing computer games for learning. *Educational Technology*, *47*(1), 20-29. http://www.jstor.org/stable/44429439

Van Mulken, S., Andre, E., & Müller, J. (1998). The persona effect: how substantial is it?. In *People and computers XIII* (pp. 53-66). Springer, London. https://doi.org/10.1007/978-1-4471-3605-7_4

Veletsianos, G., Miller, C., & Doering, A. (2009). EnALI: A research and design framework for virtual characters and pedagogical agents. *Journal of Educational Computing Research*, *41*(2), 171-194. https://doi.org/10.2190/EC.41.2.c

Vernieuwenderwijs. (2021). *Questionnaire IMI*. https://www.vernieuwenderwijs.nl/motivatie- meten-2-vragenlijsten/

Vogel, J., Vogel, D.S., Cannon-Bowers, J., Bowers, C.A., Muse, K., & Wright, M. (2006). Computer gaming and interactive simulations for learning: a meta-analysis. *Journal of Educational Computing Research*, *34*(1), 229-243. https://doi.org/10.2190/FLHV-K4WA-WPVQ-H0YM

Wiggins, G. (1998). *Educative assessment* (1st ed.). San Francisco: Jossey-Bass.

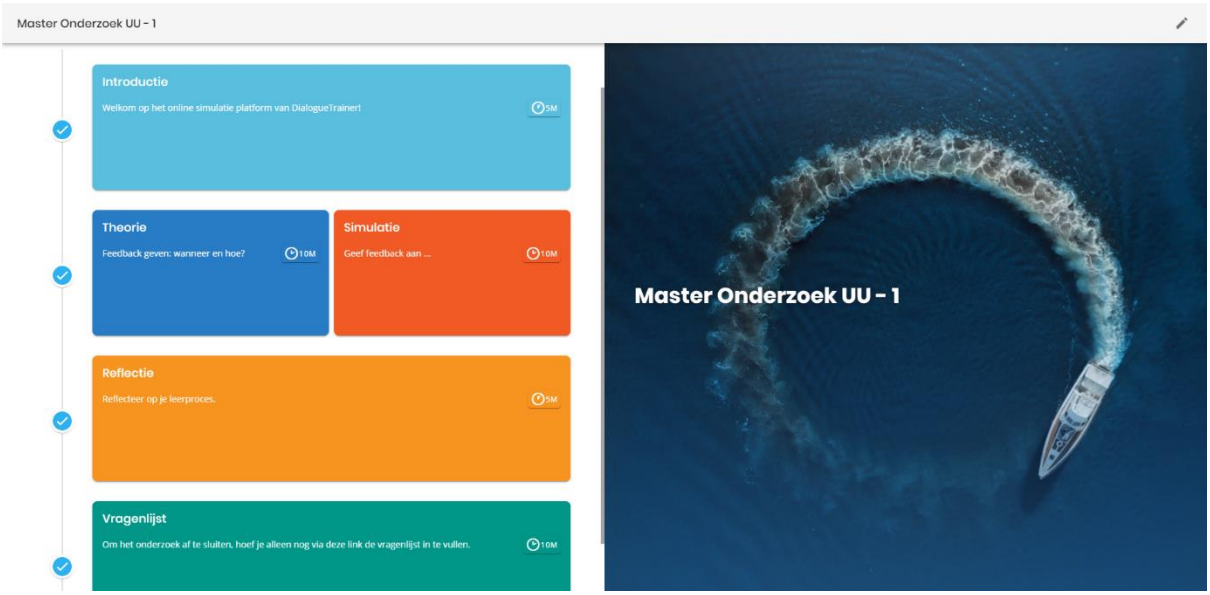
Zhonggen, Y. (2019). A meta-analysis of use of serious games in education over a decade. *International Journal of Computer Games Technology*, *2019*(1). https://doi.org/10.1155/2019/4797032

Appendix A

All participants in the experiment were directed to DialogueTrainer website and were requested to create an account there if they had not had one already. Then, depending on what link they were showed by the Qualtrics survey, they were able to access either module 1 or 2. When selected, both modules would display as depicted in Figure A1. The differently coloured tiles refer to the different parts of the module. The "Introduction" tile led the player to contextual information, which was given by text in the control group, and by a virtual coach in the other condition (see Appendix B).

Figure A1

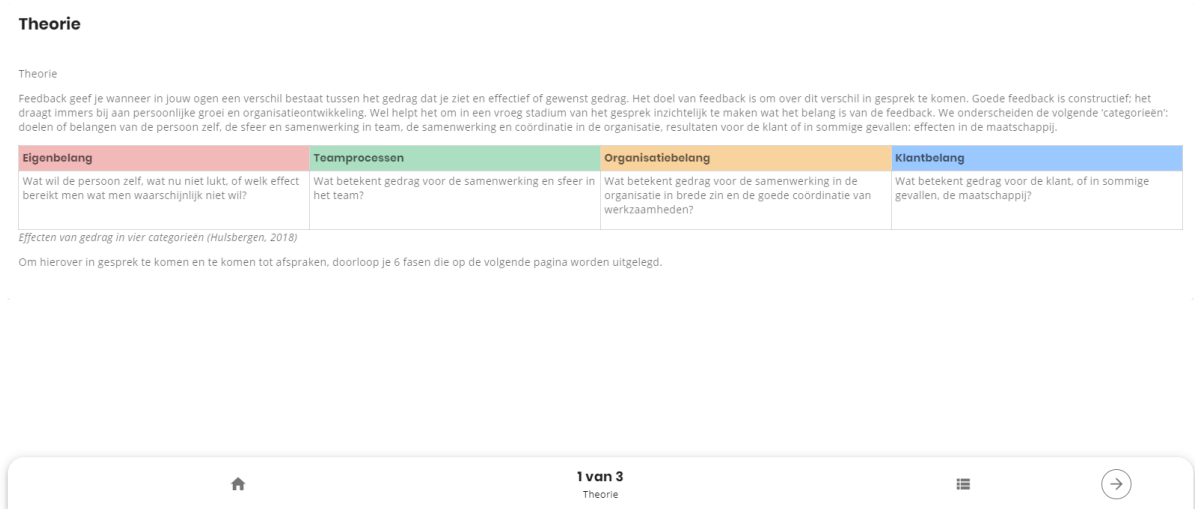
Module Display on DialogueTrainer Platform



The "Theory" tile led the player to the theory underlying the proper way in which to give feedback. As much of this information is given in the form of tables, this part of the module was kept the same in both conditions (Figure A2).

Figure A2

Example Slide Shown to Players in Theory Section of Module



In both conditions, the “Simulation” tile led to the simulation exercise. The aim of the simulation was for the player to give feedback on the performance of a colleague named Armand (Figure A3). While doing this it was important to take into account Armand’s emotions and opinions on the matter.

Figure A3

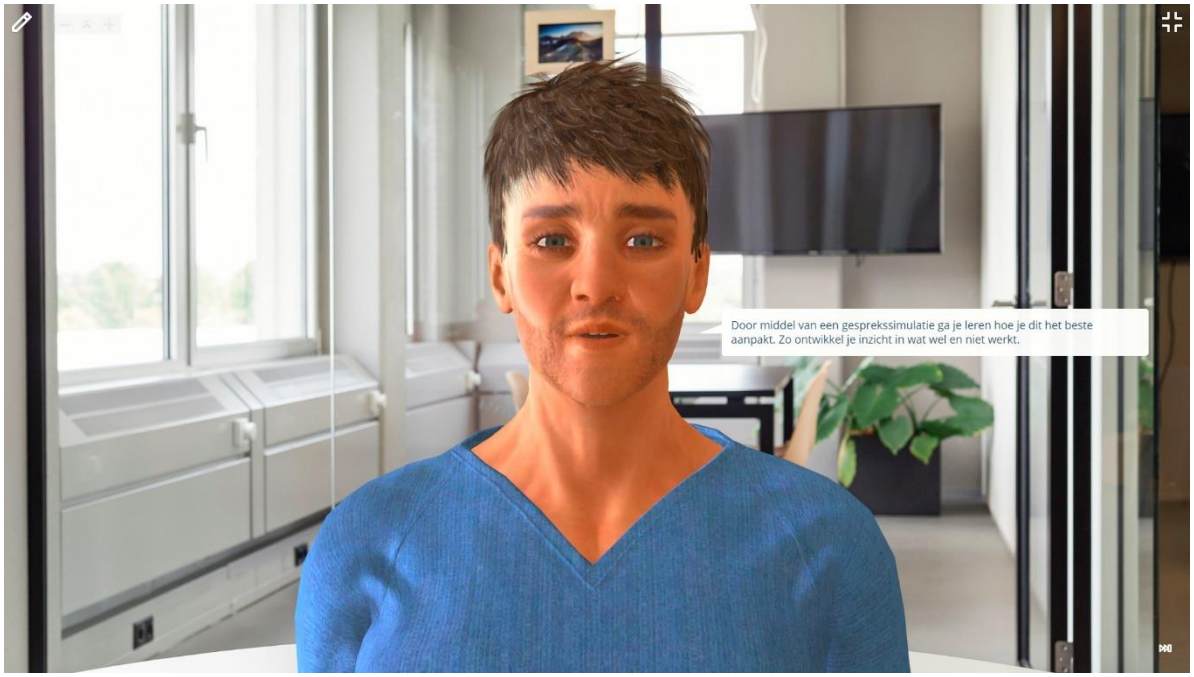
Player View During Simulation Exercise ‘Giving Feedback’

Appendix B

Between the two conditions, the only parts of the module that were different were the introduction and reflection. For the virtual coach group, selecting these tiles to a simulated interaction with Erik (Figure B1).

Figure B1

Player View During Interaction with Virtual Coach



For the control group, these tiles directed the player to slides of text. In the introduction, these contained contextual information on the use of giving feedback correctly, accounting for the other's emotions, and the role that the player can assume during the conversation (Figure B2). The reflection tile consisted of a simple reflection exercise (Figure B3). The information found in these slides has been copied almost exactly into the script of the virtual coach depicted above. By doing this, neither group would be benefited over the other regarding the amount of information they were provided. If one group was provided more information, this could result in higher scores during the simulation exercise. What was added to the virtual coach, however, were lines which aimed to establish a relationship between the player and Erik and motivate and engage the player. For example, during the introduction the virtual coach was given lines such as "Hello there, I'm Erik! How are you?" and "Are you excited to try it out?" after supplying the player with information. During the reflection exercise, the virtual coach was able to react to how well the player thought it went

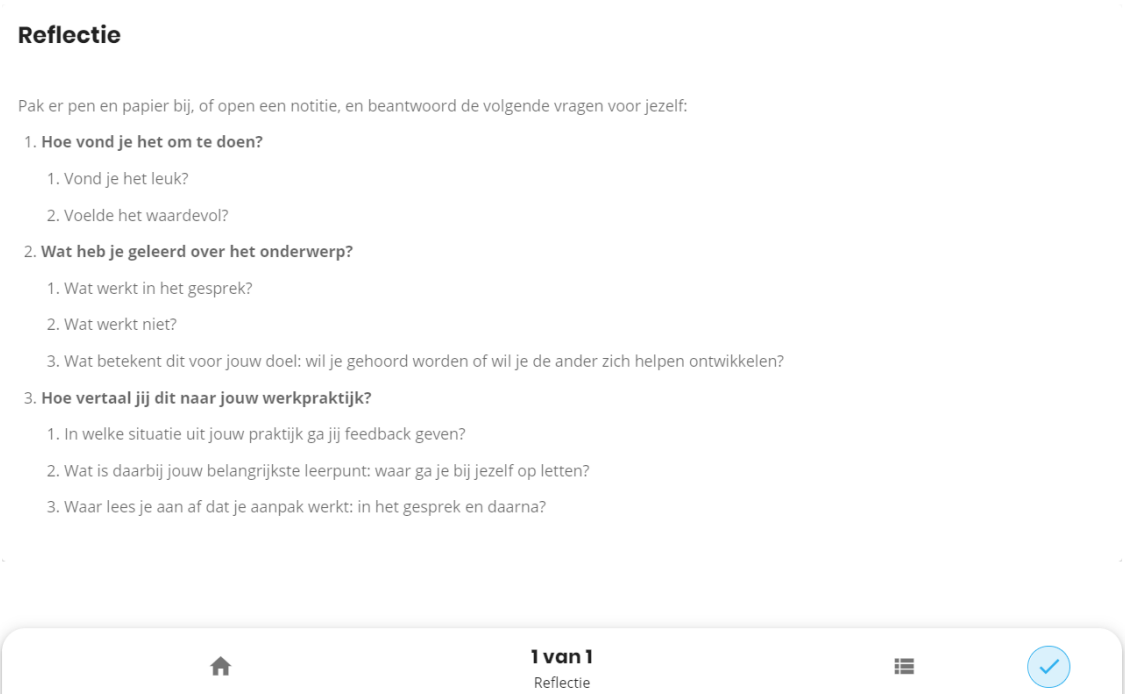
with lines such as “I’m sorry it didn’t go well but giving feedback correctly can be complicated” or “Well done! I hoped you like doing it.”

Figure B2

Example Slide Shown to Players in Introduction Section of Module

Figure B3

Example Slide Shown to Players in Reflection Section of Module



Appendix C

Questionnaires used to measure player experience

Intrinsic Motivation Inventory after Ryan & Deci, 2000.

1. Ik vond deze activiteit erg leuk om te doen.
2. Ik voelde mij niet nerveus terwijl ik bezig was met de activiteit.
3. Ik had voor mijn gevoel de keuze om de activiteit wel of niet te doen.
4. Ik denk dat ik best wel goed ben in deze activiteit.
5. Ik vond deze activiteit erg interessant.
6. Ik voelde mij gespannen tijdens de activiteit.
7. Ik denk dat ik best wel goed ben in deze activiteit, vergeleken met andere deelnemers.
8. Het was een leuke activiteit om te doen
9. Ik was ontspannen tijdens de activiteit.
10. Ik vond het erg leuk om deze activiteit te doen.
11. Ik had niet echt een keuze om de activiteit wel of niet te doen.
12. Ik ben tevreden met hoe ik het heb gedaan bij deze activiteit.
13. Ik was nerveus tijdens de activiteit.
14. Ik vond de activiteit erg saai.
15. Ik heb het gevoel dat ik deed wat ik wilde doen terwijl bezig was met de activiteit.
16. Ik voelde mij competent bij deze activiteit.
17. Ik vond de activiteit erg interessant.
18. Ik ervaarde druk tijdens de activiteit.
19. Ik heb het gevoel dat ik de activiteit moest doen.
20. Ik zou de activiteit als ‘erg leuk’ omschrijven.
21. Ik deed de activiteit omdat ik geen keuze had.
22. Nadat ik enige tijd bezig was met deze activiteit, voelde ik mij best competent.

SF-User Engagement Scale after O’Brien et al., 2018.

1. Ik verloor mezelf in deze ervaring.
2. Ik was verdiept in deze ervaring.
3. De tijd die ik spendeerde aan het spelen van de simulatie, vloog voorbij.
4. Ik voelde me gefrustreerd tijdens het spelen van de simulatie.
5. Ik vond het spelen van de simulatie verwarrend.

6. Het spelen van deze simulatie vergde veel van me.
7. De simulatie was aantrekkelijk.
8. De simulatie zag er goed uit.
9. De simulatie stimuleerde mijn zintuigen.
10. Het spelen van deze simulatie was de moeite waard.
11. Mijn ervaring was lonend.
12. Ik was geïnteresseerd in deze ervaring.