Universiteit Utrecht
Master Toegepaste Cognitieve Psychologie

A comparison study between 2 VR interviews and a real interview

27.5 ECTS

Yoshi Martodihardjo
3699110
17-06-2016

Abstract

A Dutch tech company, Ordina, was developing a VR training application for job interviewees. By order of that company, we were firstly requested to find the best medium to present the application, a head-mounted display (HMD) or a Desktop monitor. The best medium, in this case, elicits the highest perceived immersion. To find the best medium we conducted a job interview under three different conditions during the development of the application. The three conditions were the Desktop condition, the HMD condition, and the real life control condition. The first factor that differed across all conditions was the presentation method. To measure the perceived immersion of the participants we used the Slater-Usoh-Steed Questionnaire and the Presence Questionnaire. We expected that the Real condition had the highest perceived immersion, followed by the HMD condition with the Desktop condition being the lowest. In contrast to our expectation, the results from the questionnaires showed that participants had an equally high perceived immersion in the Real and HMD condition. However, as expected, it also revealed that the Desktop condition had the lowest perceived immersion. Secondly, the company requested that we create a conversational partner who was able to elicit human-like verbal behavior in participants. To see if our conversational partner could achieve this, we analyzed the verbal behavior from the job interviews mentioned above. The second factor that differed was the conversational partner. This partner was an embodied conversational agent in the HMD and Desktop version or a real human in the real version. To measure the verbal behavior of the participants, we used the participants' total words spoken, words longer than six letters, disfluencies, and types. We expected the highest scores on all verbal behavior variables for the real interview. Besides, we expected the HMD and Desktop to have the same scores on the verbal behavior variables. The results from the verbal behavior variables showed that the Real condition indeed had the highest scores. As expected, the results also showed that the Desktop and HMD versions had the same scores. We advise investing most resources in developing the conversational agent with the primary focus on natural language processing. The reason for this is that the verbal behavior variables indicated that participants in both the HMD and Desktop version did not behave similarly to the Real condition. Participants' behavior might resemble real conversations more by developing the natural language processing of the application.

Maarten van der Smagt

Ordina
Richard van Tilborg

Sjoerd Stuit

1. Introduction

Recent years has shown a great increase in consumer-oriented head-mounted displays (HMD) (Lamkin, 2016). An HMD is a display which users wear on their heads (Shibata, 2002). Examples of such consumer-oriented HMDs are the Oculus Rift, Google Cardboard and Samsung Gear VR (Lamkin, 2016). A Virtual Reality (VR) is often presented using an HMD. A VR is either an actual or synthetic environment (Steuer, 1992). Usually, VR is utilized in entertainment settings such as gaming (Lamkin, 2016). Many researchers and developers have tested the usefulness of VR in other contexts such as therapy, training, and education. Besides this, there is a rise in the use of artificial intelligence (AI) in the automation of tasks (Taub, 2015). This development is accompanied by embodied conversational agents (ECAs) (Bickmore et al., 2011; Kopp et al., 2005; Swartout et al., 2010 Lane et al., 2011). ECAs are autonomous agents with a human-like body and communicative abilities. In concert with AI, they are used to automate tasks in both the digital and the real world (Bickmore et al., 2011; Kopp et al., 2005; Swartout et al., 2010 Lane et al., 2011).

1.1 Examples of ECA utilized in the real world and research

An example of where ECAs have been used to automate tasks is in museums. Herein ECAs have functioned as guides to engage with visitors and provide information where needed (Kopp et al., 2005; Bickmore et al., 2008; Bickmore et al., 2011; Swartout et al., 2010). Besides such practical implementations, several researchers have conducted studies concerning human-ECA interactions. These studies indicate that ECAs can elicit behavior in humans similar to human-human interactions (Kramer et al., Gratch et al., 2007; Gratch & Marsells, 2004). ECAs can use multiple modalities, such as speech, facial expressions and hand gestures to communicate messages. Therefore, it is possible that such an agent exhibits human-like behavior. Users can then attribute human characteristics to the agent; this process is called anthropomorphization (Cassell, 2000). Other findings suggest that users apply 'social heuristics' during human-ECA interactions. This suggestion means that they behave similarly to normal human-human social interactions. The anthropomorphization of the ECA mentioned above probably causes such behavior (Rickenberg & Reeves, 2000). Besides, people have a tendency to ascribe mental characteristics to other entities. People also have this tendency when they are not sure if the other has a mind (Caporeal, 1986). People show this same behavior when interacting with ECAs (Caporeal, 1986).

1.2 Research on Virtual Reality and IntakeVR

Apart from the use and study of ECAs, VR has also been studied in a multitude of settings. Among these settings are therapy, training, education and task performance. VR training bases itself on the assumption that knowledge or skills acquired in a VR will transfer to the real world (Waller, 1998). As stated before, training and therapy have been a focus of VR research. Studies have shown that VR therapy has positive effects on social phobias (Klinger et al., 2005; Anderson, 2005). Multiple experiments also showed that VR training could enhance social skills (Parsons, 2002, 2004).

By the research mentioned above, it seems fair to suggest that the use of VR has found success in several fields. One other field that could profit from further research is the field of conversations and interviews in VR as studies on this subject are scarce (Villani et al., 2012). Such research is relevant as people with interview anxiety may be helped by undergoing VR exposure therapy. Researchers have already proposed VR as a method for exposure therapy (Gorini and Riva, 2008). Also, researchers have used VR to treat anxiety disorders successfully (Schultheis et al., 2002; Krijn et al., 2004; Repetto and Riva, 2011; Anderson, 2005). Furthermore, research in the field of VR interviews is relevant as interview and conversation skills learned in VR could transfer to real world interviews. Studies have already shown such transfer of skills in navigation training and other fields (Rose et al., 2010; Hamblin, 2005; Waller, 1998). It is thus relevant to see if such transference also holds for job interviews in VR and if it is usable as a treatment method for job interview anxiety. Also, explorative research is needed to see whether the desktop or HMD performs better in interviews. The developers need this information to implement the correct presentation method for the application.

The Dutch tech company Ordina found success with their VR presentation training app, APPlause. Following this, they wanted to create a similar application to improve job interview skills of interviewees. IntakeVR, the name of the application, is naturally aimed at people who want to train their job interview skills. The application allows users to have a certain flexibility conventional training does not offer. For example, as it is a digital application, one can do it in the comfort of their home. Moreover, if it were mobile, one can use it whenever or wherever one sees fit. Also, such a digital application can be updated to add capabilities to the software.

Some people might be more nervous or anxious than others during a job interview (Posthuma, Morgeson & Campion, 2002). IntakeVR might reduce such anxiousness by serving as a tool for VR exposure therapy. Such a reduction has already been shown in the field of public speaking and might also be true for the area of job interviews (Anderson, 2005). Next to reducing anxiety, IntakeVR has many other potential uses. For example, it could be an extension of a coach's standard training. Also, IntakeVR could function as a refresher or test to see if users have the needed

skills. Companies could use IntakeVR as a cheaper alternative to expensive real life coaching by training more people at the same time.

1.3.1 Theory of telepresence

Researching relevant theoretical concepts could benefit the development process of this application. Therefore, in the next paragraph, we will explain telepresence and presence. These are two crucial concepts needed to fully understand the impact of VR on the user experience and behavior. Before we go into telepresence, we will first explain presence. Presence is described as the impression of being in a certain environment at a certain time (Gibson, 1966). Presence is taken for granted in an unmediated situation. 'What is there to experience for a person other than the immediate physical surroundings?' This experience is different from when a communication device mediates perception. One is then forced to perceive two independent environments concurrently: The physical environment where one is present and the environment presented by the medium (Steuer, 1992). 'Telepresence' can be defined as the experience of presence in a 'mediated environment'. Thus favoring it over the physical environment at that moment (Steuer, 1992). Telepresence is a conjunction of immersion and interactivity. It is a function of the vividness of the representation – which leads to immersion – and of the ability to interact with the virtual environment (Steuer 1992).

1.3.2 factors of immersion

The first factor of telepresence is immersion. The factors of immersion are closely related to the factors of realism because they depend on vividness. These factors are sensory breadth and sensory depth. Sensory breadth is the ability of a communication medium to present information across the senses. Gibson (1966) defines relevant perceptual systems: the orientation system, the visual system, the auditory system, the haptic touch system and the sense of taste and smell.

Sensory depth is the depth of the sensory information available for each perceptual channel. It might be described as the quality of the virtual environment. An image is perceived as being of higher quality when the informational depth is greater (Slater, 2003). Amongst the visual aspects of sensory depth are the field of view(FOV), the field of regard(FOR), display size, display resolution, stereoscopy, head-based rendering (produced by head tracking), the realism of lighting, frame rate and refresh rate (Slater, 2003). For an in-depth explanation of these aspects see Slater (2003).

1.3.3 factors of interactivity

The second factor of telepresence is interactivity. Interactivity is defined as the extent to which users can participate in modifying the form and content of a mediated environment in real time. For the purpose of this study, interactivity will be defined as a stimulus-driven variable. It is determined by

the technological structure of the medium. Multiple factors contribute to interactivity; examples are speed, range, and mapping. For an in-depth explanation of these factors see Steuer (1992).

In this study, the actual conversation is another aspect of interactivity. Research has shown that participants behave differently in human-computer conversations compared to human-human conversations (Hauptmann & Rudnicky, 1988). Participants use a smaller vocabulary, shorter answers, and fewer disfluencies when talking to a computer (Hauptmann & Rudnicky, 1988). Another study done by Hill and colleagues (2015) showed that during human-human conversations, participants used longer words and more types than when they conversed with a chatbot. The number of types is defined as the number of unique words used i.e. size of vocabulary.

1.4 The best presentation method and use of ECA

Before we return to the current study on job interviews in VR, we will first explain our definition of the 'best presentation method'. Also, we will explain why we want the human-ECA interaction to be similar to human-human interaction. Ordina wants IntakeVR to create a realistic experience for the user. The relevance of a realistic experience in a VR training application is that training transfer to real-world settings benefits from a realistic simulation (IJsselstein & Riva, 2003). The best presentation method can present the most realistic experience for the user i.e. the highest perceived immersion. We will be investigating two presentation methods that might be able to achieve this: an HMD and a desktop computer monitor. Two reasons led us to try both methods of presentation. The first is that performance in VR differs based on the presentation medium; HMD or Desktop monitor (Waller, 1998; Ruddle et al., 1999; Silva, 2009; Pausch, 1997; Gruchalla, 2004). The second reason is that a study by Sadagic and colleagues (2000) found that HMD users did not report a higher overall sense of immersion compared to desktop users.Ordina also wants IntakeVR to elicit behavior that is similar to human-human behavior to have the best training transfer to the real world. This means that the interaction between the user and the ECA needs to resemble human-human interactions (Lortie & Guitton, 2011). Such a realistic interaction - in this case the conversation - might be achieved by using an ECA. The reason that ECAs might achieve this is that humans use social behavior in their interactions with ECAs. This social behavior resembles interactions with other humans (Rickenberg & Reeves, 2000). This resemblance might also extend to speech behavior, and this is the main reason for the use of an ECA in this study. A secondary reason is that the preceding studies only focused on verbal behavior between humans and bodiless agents (Hauptmann & Rudnicky, 1988; Hill et al., 2015). Computers and chatbots have not been able to elicit behavior similar to human-human conversations (Hauptmann & Rudnicky, 1988; Hill et al., 2015). Thus adding an ECA might change a user's behavior.

1.5 The present study

The first focus of this study will be on finding the best presentation method for the application. To test this, we will compare a job interview conducted in three environments. These environments are a Desktop VR (Desktop condition), an HMD VR (HMD condition) and a real life interview (Real condition). The real life interview will serve as the control condition. This way we can see if the HMD or Desktop performs better regarding perceived immersion. We can then conclude which is the best presentation method in the context of interview training.

We will find the best presentation method by comparing the perceived immersion between conditions. To measure the perceived immersion, we will use the Presence Questionnaire (PQ) constructed by Witmer and Singer (1988) along with the Slater-Usoh-Steed (SUS) questionnaire (Slater, Steed, McCarthy & Maringelli, 1998; Usoh, et al., 1999).

In this first analysis, the perceived immersion is the dependent variable. The independent variable is the presentation method i.e. the real, HMD or desktop interview. We expect the real interview to be most immersive. This due to (tele)presence being taken for granted in an unmediated environment (Steuer, 1992). We expect participants to have a higher perceived immersion in the HMD experiment compared to the Desktop version. This is because a more sensory breadth and a greater sensory depth would be closer to the real world regarding sensory aspects. In turn, this would influence the perceived immersion (Bowman & McMahan, 2007). The first research question is: 'What is the best presentation method in the context of VR job interviews?'.

The second focus of this study will be the verbal behavior of the participants caused by the ECA. To investigate this, we will compare the results from two human-ECA interviews with one human-human interview. This way we can see if participants' verbal behavior from the human-ECA interviews resembles the participants' verbal behavior from the human-human interview. To measure verbal behavior, we will use the total word count, disfluencies, total words longer than six letters, and total types, which were used by in the studies done by Hauptmann & Rudnicky (1988) and Hill and colleagues (2015).

In this second analysis, the verbal behavior is the dependent variable. The independent variable is the conversational partner i.e. the ECA or the human interviewer. We expect the participants to use fewer words, shorter words, fewer disfluencies, and fewer types in both the desktop and the HMD experiment compared to the real interview. This expectation is in line with the studies done by Hauptmann & Rudnicky (1988) and Hill et al. (2015). The second sub-question is: 'Can the ECA elicit verbal behavior that is similar to human-human verbal behavior?'.

Taken together, the present study will firstly attempt to find most immersive presentation method using questionnaires. Secondly, this study will attempt to discover whether the ECA can

elicit verbal behavior that is similar to human-human verbal behavior. This will be done by analyzing the verbal behavior variables. These two pieces of information will then be used during the development of the job interview training application, IntakeVR.

2. Method

2.1 Design and Procedure

This study consists of a job interview in Dutch and three questionnaires. The experiment simulated an interview concerning an IT position focusing on general software engineering. The participants sat at a desk in a room with a computer. They wore a head-mounted display (HMD) in the first condition. In the second condition, they sat across a computer monitor. In the 3rd condition, the participants sat in a room with an interviewer. There were two versions of the VR interview, an HMD version, and a desktop monitor version. The interview with the real interviewer is the 'Real condition'.

Each participant sat in a comfortable chair. The researcher explained the experiment to the participant. The participant then signed the informed consent and proceeded with the Immersion Tendency Questionnaire (ITQ). We used the ITQ to indicate if immersive tendencies differed between conditions. In the HMD condition, the HMD was placed on the participant. In Desktop condition, the screen simply turned on, and the application started. In the Real condition, the participant sat across a desk from the interviewer. At this point, the interview would begin. It lasted for about 3 to 6 minutes depending on the answer length of the participants. The participant had to use the spacebar to record his or her verbal response in both the HMD and Desktop condition. The participant was instructed to press and hold the spacebar like a walkie-talkie. After recording their message, the application processed the answer and gave a reaction. This process went back and forth until the end of the interview. All recorded audio was saved and transcribed. After the experiment, we compared the audio file to the transcript to correct the transcript where needed. We later used the transcript to extract the verbal behavior variables. The experiment closed with two questionnaires, the Presence Questionnaire (PQ) and the Slater-Usoh-Steed Questionnaire (SUS). All questionnaires were completed using a desktop computer, the monitor, and Google Forms.

2.2.1 Materials

To present the virtual application and the ECA, we used an Asus ROG desktop computer. The computer ran Windows 10. It had an i7-5820K processor, 32 GB of RAM and 2 NVIDIA GeForce GTX 970 graphics cards. For the presentation of the sound, we used the Hercules XPS 2.040 Slim. To record the voice of the participant, we used the microphone of a Logitech C270 webcam. We used

Unity 5.3.4f1 to both, run and develop the experiment. To produce the mouth movement, we used Salsa by Crazy Minnow Studio. We used Adobe Mixamo to generate the animations. We used Adobe Fuse to create the digital body of the ECA. Textures were created using Adobe Photoshop. For the text to speech and speech to text processing, we used Dragon Nuance. We used Api.AI to process the language. We used Google forms to fill out the questionnaires by the participants. The questionnaires used were the SUS (Usoh et al. 2000), the ITQ (Witmer & Singer, 1998) and the PQ (Witmer & Singer, 2005). All questionnaires are available in Appendix A, B, and C respectively. We used ITQ to measure the immersion tendency of the participant. We used the PQ and the SUS to measure the perceived immersion. All questionnaires used a 7 point Likert-scale. IBM SPSS 20.0, Microsoft Word 2016 and Microsoft Excel 2016 were used to analyze the data. The researcher participated in a job interview training at Ordina and recorded multiple interviews. This experience was used to create questions for the interview used in this study. Also, cooperation with several coaches provided insight and concrete questions and answers. All interview questions are available in Appendix D.

2.2.2 Conversational agent (API.Ai)

To conduct the conversation (in Dutch) with the participant, we used API.Ai as a part of the ECA. API.Ai is a web service that can be used to build speech-to-text, natural language processing systems. For more information about developing using API.Ai, we refer you to their website (API & Docs, 2016).

In our case, the conversational agent was able to react based on predetermined contexts and predetermined keywords. For example, in the context: 'ervaring' (experience) The following question came up:

'Hoeveel jaar ervaring heb je in totaal met softwareontwikkeling?'

(How many years of experience do you have developing software?)

In this case, the API searches the participant's answer for a number in letter form (one, two, etc.) or digit form (1, 2, etc.). It then uses that number in its next response.

An example of a participant's answer could be: '7 jaar' (7 years). The API's response would be:

'Heb je in die 7 jaar ook met Scrum moeten werken?'

(Did you work with Scrum in those 7 years?)

2.2.3 HMD

To present the HMD condition, we used an HMD (Oculus Rift DK2). The Oculus Rift DK2 is a head mounted Virtual Reality headset. The screen size of the device was 5,7 inch. The display resolution was 1920 * 1080 pixels (960 * 1080 pixels per eye) with a contrast ratio of 1000:1. The refresh rate was 75hz with a response time of 5ms. The distance to the screen was 6.4 cm. There was a horizontal visual angle of 100 degrees' and vertical visual angle of 100 degrees'. The Oculus allowed the participants to look around in the VR. Thus the Field of Regard had a full 360 degrees' in the both horizontal and vertical angle.

2.2.4 Desktop monitor

To present the Desktop condition, we used a Benq GL2450 computer screen. The screen size was 24 inch. The display resolution was 1920 * 1080 pixels with a contrast ratio of 1000:1. The refresh rate was 75hz with a response time of 5ms. The distance to the display was 60 cm. There was a horizontal visual angle of 18 degrees' and vertical visual angle of 10.6 degrees'. These angles were also the Field of Regard in the case of the Desktop condition.

2.3 Stimuli

In the case of the Real condition, the room was an office with two chairs and a desk. The interviewer was the researcher and used the same structured interview questions as in the VR interviews. The stimuli in the VR conditions was similar, consisting of an office room with the ECA, the interview, a plant and a whiteboard. The ECA was a blonde Caucasian female avatar with a height of 175 cm. The ECA sat in a chair, greeted the participant and the interview commenced. We provide a still of both the room and the ECA below (figure 1). The Desktop condition had no possibility to look around. Therefore, the FOR and FOV were identical. The HMD condition had the ability to look around due to the head tracking of the HMD. The sensory depth (FOV and FOR) in the HMD condition is thus greater than the Desktop condition. Also, the participants in the HMD condition saw either the application or the black casing of the HMD. The participants in the Desktop condition saw the application on the computer monitor and the actual room surrounding them.

Figure 1. A still from the Desktop condition. With the female avatar, a plant and a white board. Also a start and stop button at the right top to control the software. On the right side, there is a textbox to display the processed text of the participant and another textbox to display the text spoken by the ECA.

2.4 Participants

The sample consisted of 37 Dutch participants, ranging in age from 21 to 51 years (*M* = 28.68, *SD* = 7.16). They were all software developers recruited from the Dutch tech company, Ordina. One participant was excluded due to a technical failure during the experiment. The remaining 36 had an age *M* of 28.01 years and an *SD* of 6.17; 3 were female (8.33%), 33 were male (91.67%). We were not looking for a difference based on gender. Therefore, we did not evenly divide participants based on gender. The participants were evenly distributed across the three conditions; 12 (1 female; 11 males) in the HMD condition, 12 (12 males) in the Desktop condition and 12 in the Real condition (2 females; 10 males). All participants possessed normal vision or were wearing vision prescription glasses or contact lenses.

2.5.1 Statistical Analyses of the PQ, ITQ and Verbal Behavior

To analyze pre-test tendencies, we conducted a one-way Multivariate Analysis of Variance (MANOVA) 1 X 3. The presentation method was the independent variable i.e. the HMD, Desktop and Real condition. The scores on the ITQ were the dependent variable using the provided scales (Witmer et al., 1998). To analyze the influence of the presentation method on the perceived immersion, we conducted a MANOVA 1 X 3. The presentation method was the independent variable. The scores on the PQ were the dependent variable using the provided scales (Witmer et al.,

2005). We conducted a MANOVA 1 X 2 to determine the influence of the conversational partner on the verbal behavior. The conversational partner was the independent variable i.e. the ECA, and the real interviewer. The dependent variable was the verbal behavior i.e. total words, disfluencies, types and words longer than six letters. The results of the MANOVA VB and MANOVA PQ Questionnaire are available in Appendix E.

Overall, p-values <.05 were considered statistically significant (two-tailed). In the case of Box M's test, p-values <.001 were considered statistically significant (two-tailed).

2.5.2 Data reduction and further analyses

The PQ had five scales. We ran a Principal Component Analysis (PCA) to reduce the number of scales. This analysis returned one component which was called PQSUB. Therefore, an ANCOVA was conducted to determine the effect of the presentation method on the perceived immersion after controlling for verbal behavior. The verbal behavior (VBTOTAL) was the covariate.

The SUS had six questions. We ran a PCA to reduce the number of dimensions. This analysis returned one component which was called SUSTOTAL. Therefore, an ANCOVA was conducted to determine the effect of the presentation method on the perceived immersion after controlling for the verbal behavior. The verbal behavior (VBTOTAL) was the covariate.

Verbal behavior(VB) had four sub-variables, the total number of words, words longer than six letters, disfluencies, and types. We ran a PCA to reduce the number of dimensions. This analysis returned one component which was called VBTOTAL. Therefore, two ANCOVAs were run to determine the effect of the conversational partner on the verbal behavior of the participants after controlling for the perceived immersion. The first one used the perceived immersion - according to the SUSTOTAL - as the covariate. The second one used the perceived immersion - according to the PQSUB - as a covariate. The results of the PCAs of the PQ, the SUS, and VB are available in Appendix F.

3. Results

For the purpose of legibility, all unviolated assumptions were left out of the results section. We only reported violated assumptions. As stated in the method section, the results of the MANOVAs of the PQ and the VB are available in Appendix E. Next to this, the PCAs of the PQ, SUS and VB are available in Appendix F.

3.1 Pretest immersion tendency analysis

A one-way multivariate analysis of variance (MANOVA) was run to determine differences between immersive tendency across the three presentation methods using the pretest Immersive Tendency Questionnaire. Based on Wilk's Λ it is shown that there were no differences among the conditions in the pre-test Immersive Tendency Questionnaire based on the presentation method, $F(12, 56) = 2.26$ $p = 0.49$; Wilk's $\Lambda = 0.79$, partial $\eta^2 = .11$. These results indicate that there was no difference in immersive tendencies between conditions.

3.2 Influence of the conversational partner analysis

Two one-way analyses of covariance (ANCOVA) were used to analyze the influence of the conversational partner on the verbal behavior after controlling for the perceived immersion. The first one used the scores of the SUSTOTAL as a covariate. The second one used the scores of the PQSUB as a covariate. | An ANCOVA was run to determine the effect of the conversational partner on the verbal behavior (VBTOTAL) after controlling for the perceived immersion (SUSTOTAL). The covariate - the participants' perception according to the SUSTOTAL - was not related to the participants' behavior ($F(1,32) = .03$, $p = .86$, $r = .03$). There was an effect of the conversational partner on the participants' behavior after controlling for the participants' perception ($F(2,32) = 10.86$, $p = .00$, partial $\eta^2 = .40$). Planned contrasts revealed that participants in the Real condition had higher scores on the VBTOTAL compared to both the Desktop condition ($t(32) = -3.66$, $p = .00$, $r = .54$) and the HMD condition ($t(32) = -4.23$, $p = .00$, $r = .60$). Participants from the HMD and Desktop condition thus behaved differently than those from the Real condition i.e. using a lower number of all verbal behavior variables. Figure 2 shows the average scores of the VBTOTAL and the SUSTOTAL.

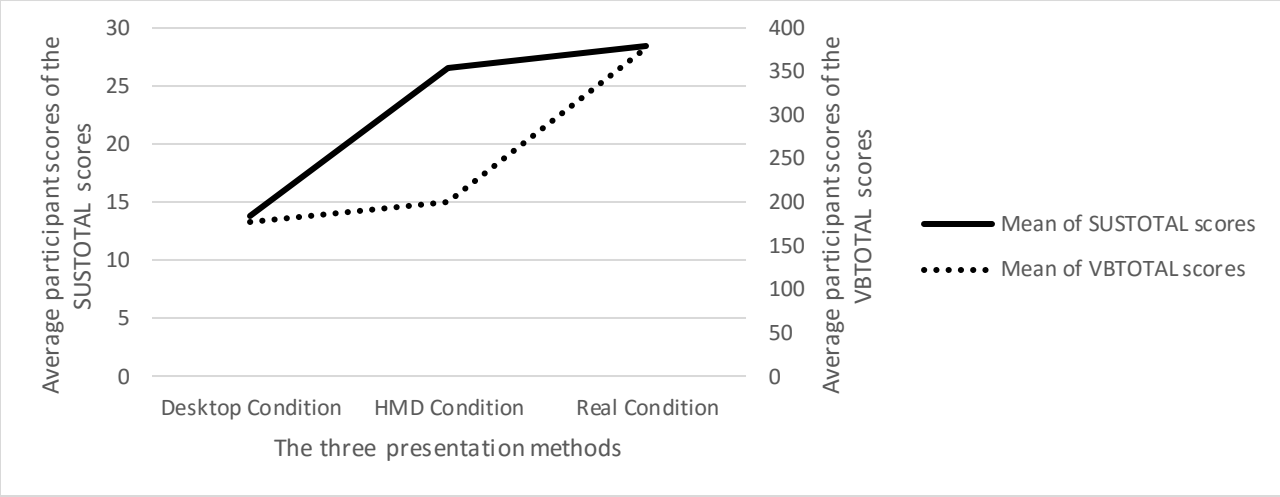

Figure 2. A graph showing the average SUSTOTAL scores on the first y-axis. The SUSTOTAL scores revealed that the participants from Real condition and HMD condition were equally high. The scores of the participants from the Desktop condition were lower than both other conditions. The second y-axis shows the average VBTOTAL scores. The VBTOTAL scores revealed that the scores of the participants of the Desktop and the HMD condition were equally low. The scores of the participants of the Real condition were higher than both other conditions.

An ANCOVA was run to determine the effect of the conversational partner on the verbal behavior (VBTOTAL) after controlling for the perceived immersion (PQSUB). The covariate - the participants' perception according to the PQSUB - was not related to the participants' behavior ($F(1,32) = .13$, $p = .72$, $r = .00$). There was an effect of the conversational partner on the participants' behavior after controlling for the participants' perception ($F(2,32) = 9.76$, $p = .00$, partial $\eta^2 = .38$). Planned contrasts revealed that participants in the Real condition had higher scores on the VBTOTAL compared to both the Desktop condition ($t(32) = -3.29$, $p = .00$, $r = .50$) and the HMD condition, ($t(32) = -4.16$, $p = .00$, $r = .59$). Participants from the HMD and Desktop condition thus behaved differently than those from the Real condition i.e. using lower number of all verbal behavior variables. Figure 3 shows the average scores of the VBTOTAL and the PQSUB.

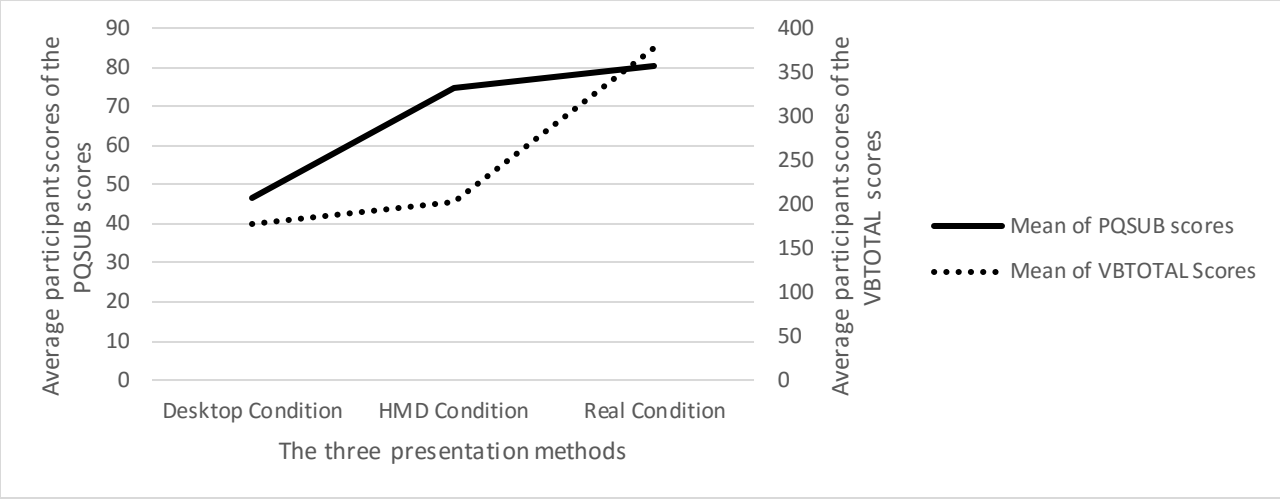

Figure 3. A graph showing the average PQSUB scores on the first y-axis. The PQSUB scores revealed that the participants of the Real condition and HMD condition were equally high. The scores of the participants of the Desktop condition were lower than both other conditions. The second y-axis shows the average VBTOTAL scores. The VBTOTAL scores revealed that the scores of the participants of the Desktop and the HMD condition were equally low. The scores of the participants of the Real condition were higher than both other conditions.

3.3 Influence of the presentation method analysis

We used two ANCOVAs to analyze the influence of the presentation method on the perceived immersion after controlling for the verbal behavior. The first one analyzed the influence of the presentation method on the scores of the SUSTOTAL. The second one analyzed the same influence on the scores of the PQSUB. Both used the scores of the VBTOTAL as the covariate.

An ANCOVA was run to determine the effect of the presentation method on the perceived immersion (SUSTOTAL) after controlling for the verbal behavior (VBTOTAL). We transformed the data of the SUSTOTAL scores as it violated the assumption of equality of variance. Levene's test resulted in $F(2,33) = 8.55$, $p = .00$ and Hartley's $F_{Max} = 4.72$ was above the critical value 4.5 both violating the assumption of equality of variance (Pearson & Hartley, 1954). After the transformation, we redid the ANCOVA to determine the effect of the presentation method on the perceived immersion (SUSTOTAL sqrt) after controlling for the verbal behavior(VBTOTAL). The covariate - the participants' verbal behavior - was not related to the participants' perceived immersion ($F(1,32) = .13$, $p = .72$, $r = .06$). There was an effect of the presentation method on the participants' perceived immersion after controlling for the participants' verbal behavior ($F(2,32) = 9.48$, $p = .00$, partial $\eta^2 = .37$). Planned contrasts revealed that participants from the Real and HMD conditions had an equal level of perceived immersion according to the PQSUB ($t(32) = -.017$, $p = .99$, $r = .60$). It also showed

that participants from the Desktop condition had a lower level of perceived immersion according to the PQSUB compared to the HMD condition ($t(32) = -3.12$, $p = .00$, $r = .54$). Figure 2 mentioned above shows the average scores of the PQSUB and the VBTOTAL. An ANCOVA was run to determine the effect of the presentation method on the perceived immersion (PQSUB) after controlling for the verbal behavior(VBTOTAL). The covariate - the participants' verbal behavior - was not related to the participants' perception ($F(1,32) = .13$ $p = .72$, $r = .00$). There was an effect of presentation method on the participants' perception after controlling for the participants' behavior ($F(2,32) = 11.73$, $p = .00$, partial $\eta^2 = .42$). Planned contrasts revealed that participants from the Real and HMD conditions had an equal level of perceived immersion according to the PQSUB ($t(32) = -.55$, $p = .59$, $r = .10$). It also showed that participants from the Desktop condition had a lower level of perceived immersion according to the PQSUB compared to the HMD condition ($t(32) = -3.84$, $p = .00$, $r = .56$). Figure 3 mentioned above shows the average scores of the PQSUB and the VBTOTAL.

4. Discussion

The first aim of this study was to find out which presentation method was the best in the context of VR job interviews. This finding could then lead to a decision on which presentation method - HMD or desktop - would be the best option for the IntakeVR application. This study firstly compared the PQ and SUS questionnaires from three conditions. The first condition was the Real condition which would conduct a real life interview. The second condition was the HMD condition. This condition used an HMD, a keyboard, and a VR to carry out the digital version of the interview. The last condition was the Desktop condition. This condition used a computer monitor, a keyboard, and a VR to conduct the digital version of the interview. The second aim was to create an ECA that could elicit verbal behavior that resembles human-human verbal behavior. This study secondly compared the verbal behavior of three conditions. The first one was the human interviewer in the Real condition. This second one was the ECA in the HMD condition. The last one was the ECA in the Desktop condition.

4.1.1 Results from the pre-experiment Immersion Tendency Questionnaire

The results from the first pre-experiment questionnaire - the Immersion Tendency Questionnaire - showed that all conditions did not significantly differ from each other. This result means that participants in every condition on average did not have a higher tendency to be immersed than participants from another condition. This result leads to the conclusion that their immersive tendencies are not an explanation for the results of the other analyses.

4.1.2 The perceived immersion hypothesis

Our hypothesis concerning the participant's perception was that the Real condition had the highest perceived immersion. Next to this we expected that the participants from the HMD condition had a lower perceived immersion, and the participants from the Desktop condition had the lowest perceived immersion. We must reject our hypothesis as we have found that the participants from the Real and HMD condition have an equally high perceived immersion on average. This result is not in line with the literature; the Real condition was not more immersive than the HMD condition which should be the case according to Steuer (1992). A reason for this result could be that participants only experienced one condition. Therefore, we do not know anything about the within-participant perception of the different conditions. This lack of knowledge in concert with the questionnaires might have led to the high scores in the HMD condition. The questionnaires in turn globally asked: 'was this experience immersive?' instead of 'was this experience immersive compared to the real world?'. The general answer in both the HMD and Real condition was 'yes, this

experience was immersive'. The reason for this response in the HMD condition was probably due to the sufficiently great sensory depth in this condition. However, we also found that the participants from the Desktop condition had the lowest perceived immersion. This last result is in line with results from Bowman & McMahan (2007). The difference in sensory depth is probably the cause of this difference in perceived immersion; the HMD condition had a greater sensory depth which in turn produced a higher level of perceived immersion.

4.1.3 Verbal Behavior Hypothesis

Our hypothesis concerning the Verbal Behavior stated that the Real condition had the highest scores on all four variables compared to the HMD and Desktop condition. These variables were word count, words longer than six letters, disfluencies, and total types. Our results support this hypothesis. The Real condition had the highest scores on all variables and the HMD, and Desktop conditions had the same scores. We thus fail to reject our hypothesis. This result is in line with the results from both Hill and colleagues (2015) and Hauptmann and Rudnicky (1988). The reason for these results is probably due to the participants knowing who they were dealing with. In the case of the Real condition, participants knew they were conversing with another human. Therefore, they probably expected an adult natural language processing. However, participants knew they were talking to a computer in the case of the HMD and Desktop condition. The expectation of an adult level of natural language processing might have lacked in the case of these conditions. Thus, they expected a lower comprehension and lower language processing skills. Therefore, they used fewer words, shorter words, and fewer types. Also, the tools - the microphone and the web services - might have caused the application to request a repeat of the participant's last answer. This repetition might have caused the participant to use fewer words, shorter words, and fewer types during the follow-up answers. Disfluencies were more frequent in the Real condition compared to both the HMD and Desktop condition. Disfluencies are a common occurrence in human-to-human interactions. These were therefore expected in the real interview (Bortfeld et al., 2001). Both the HMD condition and Desktop condition had to communicate using the spacebar. It is possible that this method gave the participants time to answer fluently. It might have even given them time to practice their answer causing fewer disfluencies.

4.2 Overall Conclusion and implications

Overall, it seems that people perceive a higher level of immersion in the HMD condition compared to the Desktop condition. This level of immersion resembles real world experiences according to the questionnaires. Thus the HMD would be the best presentation method, answering our first question. However, the participants did not behave like they were talking to a real person in both virtual

worlds. They might only have had a sense of telepresence regarding visual immersion. They did not have this sense of telepresence concerning their interactions with the ECA. Answering our second question, this result tells us that an ECA was not enough to elicit verbal behavior similar to human-human interaction in our case. A higher ability of natural language processing is probably needed to elicit this behavior. Further analysis showed that participants' perception did not influence their behavior and vice versa. This means that an increase in behavior scores would not show an increase in sensory-perceptual scores and vice versa. In practice, this ensures us that the visual aspects of the application can probably be developed independently from the level of interactivity and vice versa. One can thus safely add more features to the ECA, i.e. more natural language processing capabilities, hoping to elicit a more human-human like interaction. One would in that case probably not have to worry about a change in sensory perception.

4.3 Problems and errors

During our study, we came across several problems and errors. There was one case where the application failed to save the data (technical error) which is less than a 3 percent fail rate. This failure was not a big problem in our case, as we found a replacement participant relatively soon. The desktop version of the application had a feedback window next to the ECA. This window functioned as a feedback screen for development purposes. The developers forgot to take it out before the experiment. In our experiment, the window showed the participants exactly what the ECA said, and it showed the participant's speech registered by the application. The application did not always register the participant's speech correctly. Therefore, it might have influenced the perception of the participants in the desktop version as they saw the processed text. Either the microphone or the speech processing service caused these inaccuracies. It is still unclear which one caused this. The lenses of the HMD had scratches on them. We were not able to fix that; this unclear view might also have influenced the results of the HMD condition.

The interview followed a standard script. Certain keywords were needed to progress the interview in the Desktop and HMD conditions. Surrounding words were not necessary. In some cases, any answer would suffice. The application would be less credible if participants became aware of these faults. Dragon Nuance provided the ECAs voice. The voice, however, did not sound very natural. We are unaware of natural sounding text-to-speech voices. This unnatural sounding voice could have influenced the credibility of the application. Lastly, the participants in both the HMD and Desktop conditions encountered delays in the conversation. Naturally, this was not present in the Real condition. Delays happened due to the language processes. For a graphical depiction of the complete process, see Appendix G. These delays could have influenced the perception of the

participants and the credibility of the application as such delays do not usually occur in the real world.

4.4 future advice

The most important advice we have is that more resources should be spent on the development of the interaction of the application. Herein, the primary focus should be on the natural language processing of the ECA. A great place to start would be the ability to process the participant's input correctly. Examples of changes that could improve this ability are: finding a better performing microphone or a better-performing speech processing service. These changes would increase the credibility as the application does not repeatedly asks what the participant said. Also, these changes would cause the application to make fewer mistakes concerning the natural language processing. In turn, the application would then be abler to return the correct output. Furthermore, the application follows a scripted interview and is therefore very static. Dynamic language processing would be relevant to create a more natural and human-like experience. Lastly, we advise to remove the text feedback screen in the Desktop version. We stated the reason for this above: the application made mistakes in the language processing, which were visible. This, in turn, made the application less credible.

4.5 Closing statement

In conclusion, only one goal is met. The HMD achieves the first goal as the perceived immersion matches that of the real world according to the questionnaires. However, the ECA fails to meet the goal of eliciting verbal behavior that resembles human-human verbal behavior. Intuitively, training should thus not transfer to real life. Therefore, we do not label IntakeVR as a useable product yet. We do acknowledge the potential of its use and with further development – primarily on the natural language processing of the ECA - we are certain it will serve its purpose.

Reference list

Anderson, P.L., Zimand, E., Hodges, L.F., Rothbaum, B.O. (2005). Cognitive behavioral therapy for public speaking anxiety using virtual reality for exposure. *Depression and anxiety 22*, 156–158.

API & Docs (2016). API.Ai Retrieved from https://docs.api.ai

Beauvois, M. H. (1994). E-talk: Attitudes and motivation in computer-assisted classroom discussion. *Computers and the Humanities*, *28*(3), 177-190.

Bickmore, T., Pfeifer, L., & Schulman, D. (2011, September). Relational agents improve engagement and learning in science museum visitors. In *Intelligent Virtual Agents* (pp. 55-67). Springer Berlin Heidelberg.

Bickmore, T. W., Pfeifer, L., Schulman, D., Perera, S., Senanayake, C., & Nazmi, I. (2008, April). Public displays of affect: deploying relational agents in public spaces. In *CHI'08 Extended Abstracts on Human Factors in Computing Systems* (pp. 3297-3302). ACM.

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and speech*, *44*(2), 123-147.

Bowman, D. A., & McMahan, R. P. (2007). Virtual reality: how much immersion is enough?. *Computer*, *40*(7), 36-43.

Cassell, J. (2000). Nudge Nudge Wink Wink: Elements of Face to Face Conversation for Embodied Conversational Agents. In: J. Cassell, S. Prevost, J. Sullivan and E. Chur-chill, eds. Embodied Conversational Agents, Cambridge, *The MIT Press* 1-20

Cipresso, P., La Paglia, F., La Cascia, C., Riva, G., Albani, G., & La Barbera, D. (2013). Break in volition: a virtual reality study in patients with obsessive-compulsive disorder. *Experimental brain research*, *229*(3), 443-449.

Caporeal, L.R. (1986). Anthropomorphism and Mechanomorphism: Two Faces of the Human Machine. *Computers in Human Behavior 2* 215-234.

Dautenhahn K., Ogden B., Quick T. (2002). From embodied to socially embedded agents: Implications for interaction-aware robots. *Cogn Syst Res* 3: 397–428.

Draper, J. V., Kaber, D. B., & Usher, J. M. (1999). Speculations on the value of telepresence. *CyberPsychology & Behavior*, *2*(4), 349-362.

French R.M. (2000). The Turing test: The first 50 years. *Trends Cogn Sci* 4: 115–122.

Gorini, A., & Riva, G. (2014). Virtual reality in anxiety disorders: the past and the future. *Expert Review of Neurotherapeutics*.

Gratch, J., & Marsella, S. (2004, July). Evaluating the modeling and use of emotion in virtual humans. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 1* (pp. 320-327). IEEE Computer Society.

Gratch, J., Wang, N., Okhmatovskaia, A., Lamothe, F., Morales, M., van der Werf, R. J., & Morency, L. P. (2007). Can virtual humans be more engaging than real ones?. In *Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments* (pp. 286-297). Springer Berlin Heidelberg.

Hamblin, C. J. (2005). *Transfer of training from virtual reality environments* (Doctoral dissertation, Wichita State University).

Hauptmann, A. G., & Rudnicky, A. I. (1988). Talking to computers: an empirical investigation. *International Journal of Man-Machine Studies*, *28*(6), 583-604.

Jo, H.J., Ku, J.H., Jang, D.P., Shin, M.B., Ahn, H.B., Lee, J.M., Cho, B.H., Kim, S.I. (2001). The development of the virtual reality system for the treatment of the fears of public speaking. *Studies in Health Technology and Informatics 81*, 209–211.

Klinger, E., Bouchard, S., Légeron, P., Roy, S., Lauer, F., Chemin, I., & Nugues, P. (2005). Virtual reality therapy versus cognitive behavior therapy for social phobia: A preliminary controlled study. *Cyberpsychology & behavior*, *8*(1), 76-88.

Krämer, N. C., Tietz, B., & Bente, G. (2003, September). Effects of embodied interface agents and their gestural activity. In *Intelligent virtual agents* (pp. 292-300). Springer Berlin Heidelberg.

Krijn, M., Emmelkamp, P.M.G., Olafsson, R.P., Biemond, R. (2004). Virtual reality exposure therapy of anxiety disorders: a review. Clinical Psychology Review 24, 259–281.

Lamkin, P. (2016, April 12). The best AR and VR headsets. *Wareable.* Retrieved from http://www.wareable.com/headgear/the-best-ar-and-vr-headsets

Lane, H. C., Noren, D., Auerbach, D., Birch, M., & Swartout, W. (2011, June). Intelligent tutoring goes to the museum in the big city: A pedagogical agent for informal science education. In *Artificial Intelligence in Education* (pp. 155-162). Springer Berlin Heidelberg.

Lee E.J. (2010). The more humanlike, the better? How speech type and users' cognitive style affect social responses to computers. *Comput Human Behav* 26: 665–672.

Lehmann, K. S., Ritz, J. P., Maass, H., Cakmak, H. K., Kuehnapfel, U. G., Germer, C. T., ... & Buhr, H. J. (2005). A prospective randomized study to test the transfer of basic psychomotor skills from virtual reality to physical reality in a comparable training setting. *Annals of surgery*, *241*(3), 442-449.

Lewis, M. L., & Frank, M. C. (2015). The length of words reflects their conceptual complexity.

Lortie, C. L., & Guitton, M. J. (2011). Judgment of the humanness of an interlocutor is in the eye of the beholder. *PloS one*, *6*(9), e25085.

Parsons, S., & Mitchell, P. (2002). The potential of virtual reality in social skills training for people with autistic spectrum disorders. *Journal of Intellectual Disability Research*, *46*(5), 430-443.

Parsons, S., Mitchell, P., & Leonard, A. (2004). The use and understanding of virtual environments by adolescents with autistic spectrum disorders. *Journal of Autism and Developmental disorders*, *34*(4), 449-466.

Pertaub, D.P., Slater, M., Barker, C. (2001). An experiment on fear of public speaking in virtual reality. *Studies in Health Technology and Informatics 81*, 372–378.

Posthuma, R.A., Morgeson, F.P., Campion, M.A. (2002). Beyond employment interview validity: A comprehensive narrative review of recent research trends over time. *Personnel Psychology*, *55*, 1–81.

von der Pütten A.M., Krämer N.C., Gratch J., Kang S.H. (2010). "It doesn't matter what you are!" Explaining social effects of agents and avatars. *Comput Human Behav* 26: 1641–1650.

Repetto, C., Riva, G. (2011). From virtual reality to interreality in the treatment of anxiety disorders. *Neuropsychiatry* 1, 31–43.

Rickenberg, R., Reeves, B. (2000). The Effects of Animated Characters on Anxiety, Task Performance, and Evaluations of User Interfaces. *Proceedings of CHI 2000* 1-6

Riva, G., Davide, F., & IJsselsteijn, W. A. (2003). Being There: The experience of presence in mediated environments. *Being there: Concepts, effects and measurement of user presence in synthetic environments*, *5*.

Rose, F. D., Attree, E. A., Brooks, B. M., Parslow, D. M., Penn, P. R., & Ambihaipahan, N. (1998). Transfer of training from virtual to real environments. In *2nd European Conference on Disability, Virtual Reality and Associated Technologies*.

Ruddle, R., Payne, S., Jones, D. (1999) Navigating large-scale virtual environments: what differences occur between helmet-mounted and desktop displays? *Presence Teleoperators VR* 8(2):157–168 doi:10.1162/ 105474699566143

Shibata, T. (2002). Head mounted display. *Display*, 23(1), 57-64

Slater, M. (2003, January). A Note on Presence Terminology. *Presence- Connect*. Retrieved from http://presence.cs.ucl.ac.uk/ presenceconnect/articles/Jan2003/melslaterJan27200391557/ melslaterJan27200391557.html.

Slater, M., Pertaub, D.P., Barker, C., Clark, D.M. (2006). An experimental study on fear of public speaking using a virtual environment. *Cyberpsychology & Behavior 9*, 627–633.

Slater, M., Sadagic, A., Usoh, M., Schroeder, R. (2000). Small Group Behaviour in Virtual and Real Environments: A Comparative Study, *Presence: Teleoperators and Virtual Environments*, 9(1), 37-51.

Swartout, W., Traum, D., Artstein, R., Noren, D., Debevec, P., Bronnenkant, K., Lane, C. (2010, September). Ada and Grace: Toward realistic and engaging virtual museum guides. In *Intelligent Virtual Agents* (pp. 286-300). Springer Berlin Heidelberg.

Taub, B. (2015, November 9). Report Finds Rise Of Artificial Intelligence Could Spark Mass Unemployment And Inequality, *IFLScience*. Retrieved from http://www.iflscience.com/technology/artificial-intelligence-could-cause-mass-unemployment-and-inequality/

Turing A.M. (1950). Computing machinery and intelligence. *Mind* 59: 433–460.

Usoh, M., Catena, E., Arman, S., & Slater, M. (2000). Using presence questionnaires in reality. *Presence: Teleoperators and Virtual Environments*, *9*(5), 497-503.

Villani, D., Repetto, C., Cipresso, P., & Riva, G. (2012). May I experience more presence in doing the same thing in virtual reality than in reality? An answer from a simulated job interview. *Interacting with Computers*, *24*(4), 265-272.

Witmer, B. G., Jerome, C. J., & Singer, M. J. (2005). The factor structure of the presence questionnaire. *Presence*, *14*(3), 298-312.

Appendix A

SLATER-USOH-STEED QUESTIONNAIRE (SUS)

1. Please rate your sense of being in the virtual environment, on a scale of 1 to 7, where 7 represents your normal experience of being in a place.

2. To what extent were there times during the experience when the virtual environment was the reality for you?

3. When you think back to the experience, do you think of the virtual environment more as images that you saw or more as somewhere that you visited?

4. During the time of the experience, which was the strongest on the whole, your sense of being in the virtual environment or of being elsewhere?

5. Consider your memory of being in the virtual environment. How similar in terms of the structure of the memory is this to the structure of the memory of other places you have been today? By 'structure of the memory' consider things like the extent to which you have a visual memory of the virtual environment, whether that memory is in color, the extent to which the memory seems vivid or realistic, its size, location in your imagination, the extent to which it is panoramic in your imagination, and other such structural elements.

6. During the time of your experience, did you often think to yourself that you were actually in the virtual environment?

Appendix B

Immersion Tendency Questionnaire

1. Do you easily become deeply involved in movies or tv dramas?

|_____|_____|_____|_____|_____|_____|_____|
NEVER OCCASIONALLY OFTEN

2. Do you ever become so involved in a television program or book that people have problems getting your attention?

|_____|_____|_____|_____|_____|_____|_____|
NEVER OCCASIONALLY OFTEN

3. How mentally alert do you feel at the present time?

|_____|_____|_____|_____|_____|_____|_____|
NOT ALERT MODERATELY FULLY ALERT

4. Do you ever become so involved in a movie that you are not aware of things happening around you?

|_____|_____|_____|_____|_____|_____|_____|
NEVER OCCASIONALLY OFTEN

5. How frequently do you find yourself closely identifying with the characters in a story line?

|_____|_____|_____|_____|_____|_____|_____|
NEVER OCCASIONALLY OFTEN

6. Do you ever become so involved in a video game that it is as if you are inside the game rather than moving a joystick and watching the screen?

|_____|_____|_____|_____|_____|_____|_____|
NEVER OCCASIONALLY OFTEN

7. How physically fit do you feel today?

|_____|_____|_____|_____|_____|_____|_____|
NOT FIT MODERATELY FIT EXTREMELY FIT

8. How good are you at blocking out external distractions when you are involved i n something?

|_____|_____|_____|_____|_____|_____|_____|
NOT VERY GOOD SOMEWHAT GOOD VERY GOOD

9. When watching sports, do you ever become so involved in the game that you react as if you were one of the players?

|_____|_____|_____|_____|_____|_____|_____|
NEVER OCCASIONALLY OFTEN

10. Do you ever become so involved in a daydream that you are not aware of things happening around you?

|_____|_____|_____|_____|_____|_____|_____|
NEVER OCCASIONALLY OFTEN

11. Do you ever have dreams that are so real that you feel disoriented when you awake?

|_____|_____|_____|_____|_____|_____|_____|
NEVER OCCASIONALLY OFTEN

12. When playing sports, do you become so involved in the game that you lose track of time?

|_____|_____|_____|_____|_____|_____|_____|
NEVER OCCASIONALLY OFTEN

13. How well do you concentrate on enjoyable activities?

|_____|_____|_____|_____|_____|_____|_____|
NOT AT ALL MODERATELY WELL VERY WELL

14. How often do you play arcade or video games? (OFTEN should be taken to mean every day or every two days, on average.)

|_____|_____|_____|_____|_____|_____|_____|
NEVER OCCASIONALLY OFTEN

15. Have you ever gotten excited during a chase or fight scene on TV or in the movies?

|_____|_____|_____|_____|_____|_____|_____|
NEVER OCCASIONALLY OFTEN

16. Have you ever gotten scared by something happening on a TV show or in a movie?

|_____|_____|_____|_____|_____|_____|_____|
NEVER OCCASIONALLY OFTEN

17. Have you ever remained apprehensive or fearful long after watching a scary movie?

|_____|_____|_____|_____|_____|_____|_____|
NEVER OCCASIONALLY OFTEN

18. Do you ever become so involved in doing something that you lose all track of time?

|_____|_____|_____|_____|_____|_____|_____|
NEVER OCCASIONALLY OFTEN

Appendix C

Presence Questionnaire

1. How much were you able to control events?

|_____|_____|_____|_____|_____|_____|_____|

NOT AT ALL SOMEWHAT COMPLETELY

2. How responsive was the environment to actions that you initiated (or performed)?

|_____|_____|_____|_____|_____|_____|_____|

NOT RESPONSIVE MODERATELY RESPONSIVE COMPLETELY RESPONSIVE

3. How natural did your interactions with the environment seem?

|_____|_____|_____|_____|_____|_____|_____|

EXTREMELY ARTIFICIAL BORDERLINE COMPLETELY NATURAL

4. How much did the visual aspects of the environment involve you?

|_____|_____|_____|_____|_____|_____|_____|

NOT AT ALL SOMEWHAT COMPLETELY

5. How natural was the mechanism which controlled movement through the environment?

|_____|_____|_____|_____|_____|_____|_____|

EXTREMELY ARTIFICIAL BORDERLINE COMPLETELY NATURAL

6. How compelling was your sense of objects moving through space?

|_____|_____|_____|_____|_____|_____|_____|

NOT AT ALL MODERATELY COMPELLING VERY COMPELLING

7. How much did your experiences in the virtual environment seem consistent with your real world experiences?

|_____|_____|_____|_____|_____|_____|_____|

NOT CONSISTENT MODERATELY CONSISTENT VERY CONSISTENT

8. Were you able to anticipate what would happen next in response to the actions that you performed?

|_____|_____|_____|_____|_____|_____|_____|

NOT AT ALL SOMEWHAT COMPLETELY

9. How completely were you able to actively survey or search the environment using vision?

|_____|_____|_____|_____|_____|_____|_____|

NOT AT ALL SOMEWHAT COMPLETELY

10. How compelling was your sense of moving around inside the virtual environment?

|_____|_____|_____|_____|_____|_____|_____|

NOT AT ALL MODERATELY COMPELLING VERY COMPELLING

11. How closely were you able to examine objects?

|_____|_____|_____|_____|_____|_____|_____|

NOT AT ALL PRETTY CLOSELY VERY CLOSELY

12. How well could you examine objects from multiple viewpoints?

|_____|_____|_____|_____|_____|_____|_____|

NOT AT ALL SOMEWHAT EXTENSIVELY

13. How involved were you in the virtual environment experience?

|_____|_____|_____|_____|_____|_____|_____|

NOT INVOLVED MILDLY INVOLVED COMPLETELY ENGROSSED

14. How much delay did you experience between your actions and expected outcomes?

|_____|_____|_____|_____|_____|_____|_____|

NO DELAYS MODERATE DELAYS LONG DELAYS

15. How quickly did you adjust to the virtual environment experience?

|_____|_____|_____|_____|_____|_____|_____|

NOT AT ALL SLOWLY LESS THAN ONE MINUTE

16. How proficient in moving and interacting with the virtual environment did you feel at the end of the experience?

|_____|_____|_____|_____|_____|_____|_____|

NOT PROFICIENT REASONABLY PROFICIENT VERY PROFICIENT

17. How much did the visual display quality interfere or distract you from performing assigned tasks or required activities?

|_____|_____|_____|_____|_____|_____|_____|

NOT AT ALL INTERFERED SOMEWHAT PREVENTED TASK PERFORMANCE

18. How much did the control devices interfere with the performance of assigned tasks or with other activities?

|_____|_____|_____|_____|_____|_____|_____|

NOT AT ALL INTERFERED SOMEWHAT INTERFERED GREATLY

19. How well could you concentrate on the assigned tasks or required activities rather than on the mechanisms used to perform those tasks or activities?

|_____|_____|_____|_____|_____|_____|_____|

NOT AT ALL SOMEWHAT COMPLETELY

20. How much did the auditory aspects of the environment involve you?

|_____|_____|_____|_____|_____|_____|_____|

NOT AT ALL SOMEWHAT COMPLETELY

21. How well could you identify sounds?

|_____|_____|_____|_____|_____|_____|_____|

NOT AT ALL SOMEWHAT COMPLETELY

22. How well could you localize sounds?

|_____|_____|_____|_____|_____|_____|_____|

NOT AT ALL SOMEWHAT COMPLETELY

Appendix D

Responses of the agent, Ina

Hallo, ik ben Ina, wat is jouw naam?

User input

Aangenaam $namen, kan je beginnen met iets over jezelf te vertellen? Wat vind jij leuk om te doen?

User input

Oke, Zoals je weet hebben we een positie voor een softwareontwikkelaar. Wat is jouw ervaringsniveau? Hoeveel jaar ervaring heb je in totaal met software ontwikkelen?

User input

Heb je in die $nummers jaren ook met Scrum moeten werken?

User input

En hoe beviel dat?

User input

Ik kan mij wel voorstellen dat je het $bijvoeglijkErvaring vindt. Wat vind je eigenlijk leuk aan je werk als Softwareontwikkelaar?

User input

oke, dat snap ik wel. Welke rol heb jij graag in een team?

User input

Dat is mooi, we kunnen wel een $rollen gebruiken. Werk je eigenlijk liever frontend of backend?

User input

We hebben iemand hard nodig voor $frontback, dus dat komt mooi uit! Kan je misschien een voorbeeld geven van een project waar je $frontback werk deed?

User input

Ik denk dat ik nu genoeg weet. Ik wil je bedanken voor je komst, en een fijne dag wensen.

*Everything with an $ is a variable and will be used by API.AI in its response.

Appendix E

MANOVA verbal behavior and MANOVA Presence Questionnaire

For the purpose of legibility, assumptions were only reported if they were violated. All assumptions that were not violated were left out.

1.1 MANOVA Verbal behavior

A One-way multivariate analysis of variance (MANOVA) was run to determine the influence of the conversational partner on the verbal behavior. Verbal behavior had 4 variables: word count, Disfluencies, Types and Six letters. We transformed word count data using the square root transformation as it violated the assumption of equality of variance. Levene's test resulted in $F(2,33) = 5.78$, $p = .01$. and Hartley's $F_{Max} = 6.41$ was above the critical value 4.5 (Pearson & Hartley, 1954) both violating the assumption of equality of variance.

Results from the MANOVA showed that the conversational partner had an effect on all variables. The p-values were between word count SQRT ($F(2,33) = 15.56$, $p = .00$, $\eta^2 = .49$ with observed power of .99) and disfluencies ($F(2,33) = 6.42$, $p = .00$, $\eta^2 = .28$ with observed power of .88).

Planned contrasts showed that participants in the real interview had higher scores on all four variables compared to both the HMD and Desktop condition. The p-values were between Disfluencies (HMD condition, $t(32) = -2.83$, $p = .01$) and total word count (HMD condition, $t(32) = -5.17$, $p = .00$).

1.2 MANOVA Presence Questionnaire

A One-way multivariate analysis of variance (MANOVA) was run to determine differences between the scores on the presence questionnaire across the three levels of immersion. PQ had five scales: Involved/Control(IC), Adaption/Immersion(AI), Visual Fidelity(VF), Interface Quality(IQ) and Sound. We transformed the data of the VF scale using the logarithm transformation as it violated the assumption of equality of variance. Levene's test resulted in $F(2,33) = 9.87$, $p = .00$. and Hartley's $F_{Max} = 17.16$ was above the critical value 4.5 (Pearson & Hartley, 1954) both violating the assumption of equality of variance.

Results from the MANOVA showed that the level of immersion had an effect on IC, AI and VF log. The p-values were between IC ($F(2,33) = 14.79$, $p = .00$, $\eta^2 = .47$ with an observed power of .99) and AI ($F(2,33) = 9.93$, $p = .00$, $\eta^2 = .38$ with an observed power of .98). IQ showed no effect ($F(2,33) = .24$, $p = .79$, $\eta^2 = .01$ with an observed power of .08), and Sound also showed no effect ($F(2,33) = .43$, $p = .65$, $\eta^2 = .03$ with an observed power of .11).

Planned contrasts showed that participants in the real interview had the same scores on all three scales that showed an effect. The p-values were between VF ($t(32) = -.55$, $p = .59$) and IC ($t(32) = -.72$, $p = .48$). The scores from the HMD condition was higher on all three scales that showed an effect compared to the Desktop condition. The p-values were between AI ($t(32) = -4.10$, $p = .00$) and IC ($t(32) = -4.10$, $p = .00$).

Appendix F

Principal component analysis of the Presence Questionnaire, The SUS Questionnaire and the Verbal behavior scores.

1.1 Dimension reduction Presence Questionnaire

A principal component analysis (PCA) was conducted on the 3 scales of the PQ with orthogonal rotation (varimax). These scales were the IC, AI, VF as these were the scales that showed an effect. The Kaiser–Meyer–Olkin measure verified the sampling adequacy for the analysis, KMO = .76 ('good' according to Field, 2009), all 3 KMO values for individual items were > .75, which is well above the acceptable limit of .5 (Field, 2009). Bartlett's test of sphericity $\chi 2$ (10) = 69.68, p = .00, indicated that correlations between items were sufficiently large for PCA. An initial analysis was run to obtain eigenvalues for each component in the data. 1 component had eigenvalues over Kaiser's criterion of 1 and in combination explained 85.41% of the variance. The scree plot showed an inflexion that would justify retaining component 1. Visual inspection of the scree plot and Kaiser's criterion indicated that only one component should be retained. Table 5 shows the factor loadings. The items that cluster on the same components suggested that component 1 represented a general PQ perception of presence. The sum of the IC, AI, and VF scales was used in the result section. This scale was referred to as PQSUB.

	Component 1
IC	.92
AI	.93
VF	.93

Table 5. Component Matrix of the scales that showed differences across the three groups of the PQ.

1.2 Dimension reduction SUS - Questionnaire

A principal component analysis (PCA) was conducted on the 6 items of the SUS with orthogonal rotation (varimax). The Kaiser–Meyer–Olkin measure verified the sampling adequacy for the analysis, KMO = .81 ('great' according to Field, 2009), all KMO values for individual items were > .75, which is well above the acceptable limit of .5 (Field, 2009). Bartlett's test of sphericity $\chi 2$ (15) = 109.78, p = .000, indicated that correlations between items were sufficiently large for PCA. An initial analysis was run to obtain eigenvalues for each component in the data. 1 component had eigenvalues over Kaiser's criterion of 1 and in combination explained 62.20% of the variance. The

scree plot showed inflexions that would justify retaining component 1. Visual inspection of the scree plot and Kaiser's criterion indicated that only one component should be retained. Table 6 shows the factor loadings after rotation. All items clustered suggesting that the single component represented a general SUS perception of presence. The sum of all 6 items was used in the result section. This variable was referred to as SUSTOTAL.

	Component 1
SUS Q1	.88
SUS Q2	.86
SUS Q3	.85
SUS Q4	.80
SUS Q5	.66
SUS Q6	.66

Table 6. The component matrix of the six questions of the SUS questionnaire.

1.3 Dimension reduction Verbal behavior

A principal component analysis (PCA) was conducted on the 4 scales of VB with orthogonal rotation (varimax). The Kaiser–Meyer–Olkin measure verified the sampling adequacy for the analysis, KMO = .80 ('good' according to Field, 2009), all KMO values for individual items were > .71, which is well above the acceptable limit of .5 (Field, 2009). Bartlett's test of sphericity $\chi 2$ (6) = 202.45, p = .000, indicated that correlations between items were sufficiently large for PCA. An initial analysis was run to obtain eigenvalues for each component in the data. 1 component had eigenvalues over Kaiser's criterion of 1 and in combination explained 88.79% of the variance. The scree plot showed inflexions that would justify retaining component 1. Visual inspection of the scree plot and Kaiser's criterion indicated that only one component should be retained. Table 7 shows the factor loadings after rotation. All items clustered one the single component suggesting that the component represents a general Verbal behavior. The sum of all 4 items was used in the result section. This was referred to as VBTOTAL.

	Component 1
Types	.98
Six letter	.94
Disfluencies	.86
Word count	.98

Table 7. Component matrix of the variable of verbal behavior.

Appendix G

Graphical depiction of the complete speech process.

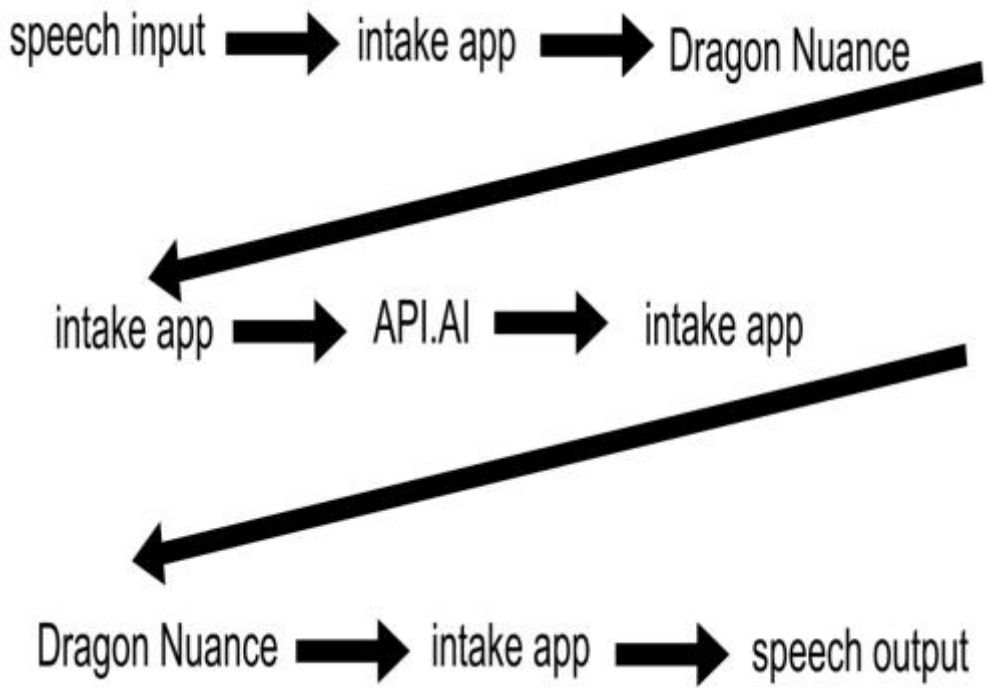

Figure 4. Graphical depiction of the process from the participants' speech until the applications speech output. The recorded speech had to be sent from the application to Dragon Nuance which converted it to a piece of text. That processed text had to be sent back to the application which sent it to Api.AI. In turn, it sent an answer back to the application based on the user's processed text. This answer would then again have to be sent to Dragon Nuance which converted the text to speech. The converted speech was sent back to the application which delivered it to the user via the speakers.