How do personality traits affect annotations of behavior in police interrogations?

Master Thesis

David Brode

S0153575
Master Human Factors and Engineering
July 26, 2014

Begeleiders:
Prof. dr. Frank van der Velde
Dr. ir. H.J.A. op den Akker
Msc Merijn Bruijnes
Universiteit Twente

UNIVERSITEIT TWENTE.

Acknowledgments

Dear reader,
The master thesis that lies before you is the product of a successful cooperation between the University of Twente, the Police Academy in Apeldoorn, The Netherlands and myself. I am, and will be, forever grateful for the cooperation, enthusiastic reactions and enduring patience of all the people involved.

Especially I would like to thank Frank, Rieks and Merijn for supervising me with patience, understanding and an open mind. Some of the ideas I had were not as straightforward as I imagined, yet they approach them with sound advice and suggestions. I would also like to express my sincerest gratitude to Ben Koster. He welcomed me with open arms to the amazing world of the Police Academy and took considerable time to show me around and explain how they work. To this day I am amazed about the capacity and thoroughness with which they operate. I would also like to thank Imke Rispens, Indre Driessen and Sabine Reurich who allowed me attendance in their classes and provided valuable insights for my thought process. I also am grateful for the cooperation and enthusiastic reactions I received from the detectives. I can imagine that a student who talks about personality, embodied conversational agents and an annotation task, was not the first thing on their mind when they signed up for a specialized course on interrogations. I owe thanks to Max Hermelink for helping me by editing the video clips that were used in this study.

I am very, very, very thankful to Juicy J, Joost Dijkstra, for his endless support in proofreading this thesis and providing me with solid ideas and hilarious pictures to keep me going. Last but not least I want to thank my mother, Steffi Brode, for supporting me always to chase my dreams and finish what I started. Being accustomed to working fulltime, it was not easy to sit down and patiently study all year, but I succeeded and hope to interest the reader with this concluding research.

Sincerely,

David Brode,

July 26, 2014

Table of contents

Abstract

This study is about personality, embodied conversational agents (ECA) and interpretation of behavior in an annotation task of police interrogations. An ECA is being developed by the University of Twente in cooperation with the police academy with the purpose of providing new ways to assist detectives learning and improving interrogations skills. This study aims to contribute to the development of a part of that system which is concerned with natural dialog and the overall realness of the interaction between a detective and the ECA. A group detectives and non-detectives completed a five factor personality test and annotation task. Leary's interpersonal theory of behavior was used to conduct the annotations. We were interested in differences between both groups on the test and task and set out to look for possible relations between the scores on the personality test and how they rated the behavior in the interrogations. Detectives scored higher than non-detectives on Agreeableness, but did not differ on the other domains. Behavior of the interviewee was rated more defiant by the group non-detectives. Overall inter-rater agreement was equal to previous findings.

Preface

Imagine you are a psychologist working for a police department, your job, amongst many other things, is to assist detectives during an interview with a suspect or witness. You are in the room next to the one where the actual interview is taking place and can see and hear everything that is happening from behind a one-way mirror connecting both rooms. You look, listen and carefully observe the detective and suspect (or witness) and provide input to the detective through a microphone (earpiece) he is wearing. You are trained and experienced in understanding human behavior and therefore realize that there is a difference between a normal conversation and a strategic interview, such as is the case with a police interrogation[1]. However convenient and valuable your input is, you are not always there. A detective on the job has to deal with many responsibilities, one of which is to conduct a strategic interview on his own. Imagine that the detective could have prepared for the interview with the use of a game-like computer program with which he could interact as if he was conducting a real interview. Virtual characters, or embodied conversational agents, in this game react to questions so that detectives can practice with different approaches, gaining more insights into behavior and how conversational strategies work.

[1] The term interrogation can also be read as interview, both are used as synonyms throughout the paper.

1. Introduction

When computer games are used for educational purposes they are known as serious games (Sawyer, 2007). Serious gaming has gained more attention in recent years due to its educational benefits and applicability in professional settings (for an overview see: Lee, Heeter, Magerko, & Medler, 2012; Gee, 2003; Gee, 2007; Prensky, 2001; Squire, 2007; Van Eck, 2006; and the *deLearyous gaming project*, n.d.). Military training programs also use virtual games to prepare soldiers for specific situations in battle (see Hill et al., 2003). Not only do serious games reduce costs and risks, the huge advantage is that users can play around and try different roles and strategies (Lee et al., 2012). But how are serious games developed, and in particular, how do you create an Embodied Conversational Agent (ECA) that is capable of natural dialog? A good introductory example is provided by the *virtual humans project* (2010) of the USC Institute for Creative Technologies. Looking into the complete development of a serious game or an ECA is beyond the scope of this paper and our study. We are mentioning these to provide a picture of the subject as a whole. Our study looks at the relation between personality and interpretation of behavior in interpersonal communication in police interrogations. This information is gathered with the purpose of aiding the development of the system of an ECA that has to deal with recognizing, reasoning and responding to stances (Bruijnes, 2013). The process of recognizing stances is fueled primarily through annotating utterances of police interrogations. When annotators agree on the meaning of the utterances, that information can be used to build the system of recognizing stances for an ECA. As it turns out agreeing on the meaning of the utterances is challenging (see also op den Akker, Bruijnes, Peters, & Krikke, 2013).

We will first explain this challenge and as might already been deduced from the previous paragraph, we conducted our experiment to discover whether there is a relation between the personality of an annotator and his interpretation of behavior of others in police interrogations using the same theoretical stance model and procedure to annotate video material as op den Akker et al. (2013) did in their study. We were allowed to temporarily join the education program on interrogations at the Police Academy in the Netherlands[2] and will describe our experience and relevance towards our study. Secondly, all personality theories used in this study will be explained and we will look at what they say about interpreting behavior of others. We then turn to our hypotheses and experiment. Next, we discuss our findings and

[2] Every reference to the Police Academy is made towards the Police Academy in Apeldoorn, The Netherlands. The teachers, policemen, detectives and organization as a whole have been very kind, helping and open-minded toward our study.

draw conclusions. We will finish with recommendations and provide suggestions for future studies on this subject.

1.1. Current challenge

The Human Media Interaction department of the University of Twente is working together with the Police Academy to build the aforementioned ECA that can assist training detectives to acquire and/or broaden interrogation skills. To accomplish this they are developing a toolbox for Learning Interaction Stances (LearIS) in a police context based on Leary's interpersonal stance theory (Bruijnes, 2013). The ECA acts as a suspect or witness and detectives can interact with it realistically (we strongly recommend reading Bruijnes, 2013, pp. 625-627 for a detailed description of the whole LearIS system). One of the challenges in building the response system comes from the annotations on which it relies in order to make sense of the input.

In the study of op den Akker et al. (2013), nine students annotated video-material of a police interview using Leary's interpersonal stance theory[3] (see op den Akker et al., 2013, p. 200). Although we will explain Leary's theory in detail later, annotating with the use of this theory means that annotators have to attribute one of the eight behavioral categories (also a ninth 'neutral' was added) to each of the utterances of the detective and, in this case, suspect. The inter-rater agreement turned out to be low (a Krippendorff's alpha was calculated, α = 0.24) when looking at each label separately and slightly improved when two labels of the same quadrant were considered equal (α = 0.42). Op den Akker et al. (2013) concluded, among other things, that "Leary's theoretical model makes sense as a framework for analyzing and describing the interactional stance people take towards each other in a social encounter" (p. 212). Difficulties arise when annotators are forced to choose from set labels because they may attribute different meanings to the words used to describe the labels.

During the aforementioned study, annotators discussed the behaviors of both the detective and suspect (after the annotation task). These discussions, which yielded broad perspectives considering there were only nine students, together with our background in psychology, lead to the question in what way the difference in interpretation of behavior (as categorized by Leary's interpersonal stance theory) is related to the personality of an annotator. The concept behind this assumption is that if there indeed appears to be a distinct relation between the personality of an annotator and the interpretation of behavior then this information can be

[3] Disclaimer: I was one of the nine students to annotate the video-material.

used to make (interaction with) an ECA more realistic, but also to better understand the differences that arise when material is being annotated. Naturally, other benefits could also be imagined, such as personalized learning goals. Time at the Police Academy strengthened the idea that there could be a difference how detectives and non-detectives interpret behavior of others.

1.2. Visiting the Police Academy

When we discussed our initial ideas about the influence of personality on the interpretation of behavior of others we were searching for theoretical underpinnings on this subject. There remained however an idea that if the current study was meant to assist in the development of a serious game for the police, their input and perspective is paramount. After we presented our ideas and experiment we were temporarily allowed attendance in a specialized course on interrogations at the Police Academy. This experience provided valuable insights, we will mention two that helped shape our ideas and hypotheses.

The first insight was the overwhelming capacity of the organization as a whole. The Police Academy is innovating and the teachers there are very open about their work, the possibilities and limitations with which they have to cope. Focusing solely on the capabilities and skills that are required to conduct an interview we quickly realized just how much potential a digital tutoring system can have (see Smets, 2011, p. 46; and Gudjonnson, 2002 for an elaborate explanation of the skills needed to conduct a police interrogation). When we explained the workings and possibilities of an ECA in a serious game setting the detectives reacted positive and could picture themselves using a serious game as long as the interaction is realistic (Detectives from the Police Academy, personal communication, April 29, 2014). Although this may sound straightforward, it is not. As we have come to understand and appreciate during the classes, detectives experience a lot of hardship and witness or come in contact with traumatic events. Offering a game to learn real life skills might seem childish and not adequate. Learning, after our explanation and examples, that they would consider using the program created mutual understanding. This was important for us because we were dependent on their cooperation for our study.

The second insight is related to the perceived difference in personality between students of the University (that had annotated the material used in the study by op den Akker et al., 2013) and the detectives we met at the Police Academy. To us this distinction was important because it fueled the notion that if the difference in personality was as clear cut as we seem to perceive it, the development of an ECA needed annotation input from detectives so that the system could be made more realistic and fine-tuned for their use. It remains to be seen if the difference in personality is indeed significant and if it will lead to different interpretation of behavior between detectives and non-detectives when observing police interviews. In the next section different personality theories will be explained to form a theoretical framework about the relation between personality and the interpretation of behavior in interactions between people.

1.3. Leary's interpersonal circumplex

Leary's (1957) and associates' interpersonal model is at the heart of our study. It is implemented as a computational model for interpersonal stance in the ECA (Bruijnes, 2013), we will use it in our study to annotate the video material and the Police Academy teaches detectives how to effectively use the model to communicate in interrogations, to understand how certain behavior emerges and how to influence that behavior (van Amelsvoort, Rispens, & Grolman, 2012). A thorough explanation is therefor in order, but first a clarification might be of use. When you read about Leary's Rose, Leary's interpersonal circumplex, Leary's interpersonal theory of behavior, the Leary framework, the interpersonal behavior circle or Leary's interpersonal stance theory, they all refer to the same model. The general notion is that all behavior we exhibit when we interact with each other can be thought of as either *dominant* (above) or *submissive* (under) and *together* or *opposed*. Within Leary's Rose (see figure 1, p. 9) these categories are displayed on two orthogonal axes (a circle is drawn around for clarity): *dominance* and *submissiveness* on a vertical axis and *opposed* and *together* on a horizontal axis. This yields four quadrants, which are then further specified with two descriptions of behavior in each quadrant resulting in a total of eight octants or, simply put, eight ways to express oneself when interacting with others.

Discussing and understanding these behaviors is easier when they are first labeled according to their position (see table 1, p. 9): *leading* behavior is in the upper right quadrant touching the vertical axis, we call this above and together, or AT. *Helping* behavior is still in the upper right quadrant, but touches the horizontal axis, we call this together and above, or TA. Continuing clockwise the next two behaviors are in the lower right quadrant. These behaviors are *cooperative*, which is together and under, or TU and *dependent* or under and together, UT. The behaviors in the lower left side are *withdrawn*, or under and opposed, UO and *defiant* or opposed and under, OU. The upper left two behaviors are *aggression*, or opposed and above, OA and *competitive* or above and opposed, AO. The abbreviations and noticing where a label touches the axis helps to better understand the subtleties between different behaviors.

The next step is to understand the proposed dynamic of the interactions. According to Leary's (1957) theory dominant behavior induces submissive behavior of the other interlocutor and vice versa (see figure 2). In the theory another distinction is made in regard to the severity of each behavioral tendency. The more someone's behavior moves to the outside of the circle the more intense and unconstructive it is (see Carson, 1969; and Van Dijk & Cremers, 2013). With the use of this model the Police Academy can teach detectives to shift an interrogation from a 'normal' conversation to a strategic interview (B. Koster, personal communication, April 29, 2014). Leary's Rose not only explains the behaviors we use when

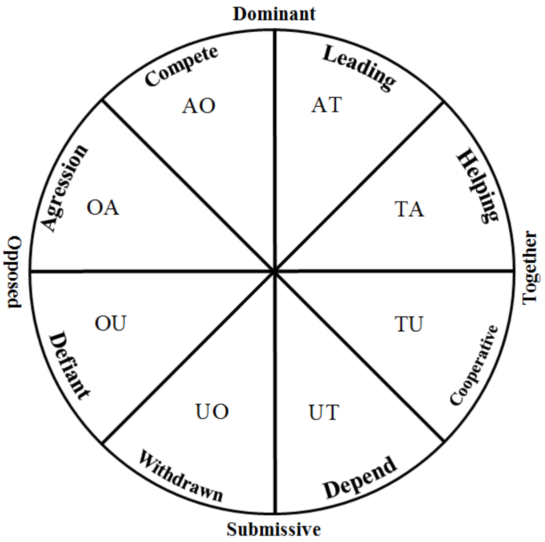

Figure 1. Leary's Rose.

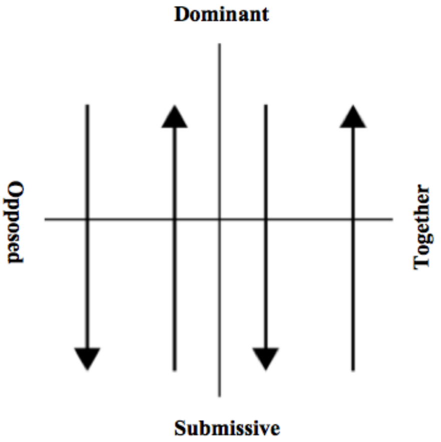

Figure 2. Leary's Rose, the arrows show the behavioral tendencies according to Leary's theory.

Table 1. Behavioral categories, examples, octants and elicited behavior

Behavior	**Example**	**Octant**	**Elicits**
Leading	"I will propose some options and we will choose from those"	AT	Depend
Helping	"I think you are right"	TA	Cooperative
Cooperative	"Tell me what I can do for you"	TU	Helping
Depend	"If you think this is the way, then we will go with that"	UT	Leading
Withdrawn	"Yeah, ok, whatever, my opinion does not matter that much"	UO	Compete
Defiant	"Yes, maybe, but if you are going to act like that, then…"	OU	Aggression
Aggression	"I absolutely do not agree with that"	OA	Defiant
Compete	"I know what's best, this is the way it should be done"	AO	Withdrawn

Note. There are many other examples possible, these were chosen to highlight the behavioral tendency using Van Dijk & Cremers (2013).

we interact with each other, the model also proposes that in our interaction we are behaving in ways that produces expected behavior in others. When for instance, someone asks for help, he expects leadership behavior from another and when he helps someone else, he expects cooperation. But why is this important to us? Carson (1969) makes another assumption that will underpin the current study:

> We would expect reasonably well-adjusted persons to be capable, in appropriate circumstances and with modulated intensity, of displaying behaviors across the entire range of the eight categories. It will usually be the case, however, that a particular person's social behavior will favor some segments of the circle more than others, thus giving his interpersonal behavior the distinctive coloration we ordinarily associate with the concept of *personality*. (p. 112)

Carson (1969) notes: "when Leary's and his associates made the initial suggestions that interpersonal behavior may be circumplicially ordered, it was not more than intuitive prescience" (p. 103). In later years much research was done and the framework proved to be robust, "despite minor variations and deviations" (Carson, 1969, p. 106) (see Roe, 1957; Borgatta, Cottrell, & Mann, 1958; Schaefer, 1959; Slater, 1962; Schaefer, & Bayley, 1963; Becker, & Krug, 1964; Lorr, & McNair, 1963, 1965; and Lorr, Bishop, & McNair, 1965). Although these many different studies reveal other interesting representations of an interpersonal circumplex, we will stick to the introduced model because it is used by the Police Academy, in the development of a system for an ECA and in our experiment.

As briefly mentioned, theory suggests that when people interact with each other there is an element of expectation towards the response they will get. In order for that to happen successfully, the interpretation of the message that was sent has to be in line with the sender's expectation. We will now turn to the subject of personality using the Five Factor Model (Costa & McCrae, 1992) and discuss what role personality plays in the interpretation of behavior, because as we mentioned, the goal of this thesis is to see if there is a distinct relation between someone's behavior and the way they interpret the behavior of others (as categorized by Leary's interpersonal stance theory) in police interrogations. If true, then this information can then be used to make recommendations towards the development of the system of an ECA.

1.4. Five Factor Model and interpretation of behavior of others

People have distinct personalities, this might be considered common knowledge, just like it is little surprising that we act differently across situations and interactions because of who we are and the experiences we have had in life. These assumptions remain valid, but we also agree that certain people seem to fit better in some places than others. We go through great trouble finding and selecting the 'right' people for a job, but also in other areas of our lives we look for specific qualities in people. Personality traits shape our personality and directly influence our behavior. We know that they are relatively stable over time (Matthews et al., 2009) and although there can be differences between behaviors we express in some situations, there are also more innate construct of who we are and how we tend to (re)act that are stable over our life span (Roberts, Walton, & Viechtbauer, 2006).

Looking at someone's personality is one way to say something about the amount of expected 'fit' in certain functions. In fact, tests relating to personality are numerous, but they are not all very reliable and some suffer more from deficiencies than others (Matthews, Daery, & Whiteman, 2009). They noted that the personality theory of Eysenck (1967, 1997) and the Five Factor Model of Costa and McCrae are "two prominent personality schemes which advocate the usefulness of higher-order secondary factors, describing personality in broad, abstract terms. Within these schemes each dimension may be assumed to be significantly related to hundreds of basic trait terms" (Matthews et al., (2009), p. 23).

Table 2. Trait facets associated with the five domains of the Costa and McCrae five factor model of personality

Neuroticism (N)	anxiety, angry, hostility, depression, self-consciousness, impulsiveness, vulnerability
Extraversion (E)	warmth, gregariousness, assertiveness, activity, excitement-seeking, positive emotions
Openness (O)	fantasy, aesthetics, feelings, actions, ideas, values
Agreeableness (A)	trust, straightforwardness, altruism, compliance, modesty, tender-mindedness
Conscientiousness (C)	competence, order, dutifulness, achievement, striving, self-discipline, deliberation

Adapted from "Personality traits" by G. Matthews, I. J. Daery and M. C. Whiteman, 2009, Cambridge: University Press, p. 25. Copyright 2009 by G. Matthews, I. J. Daery and M. C. Whiteman.

We focus on the latter model, which is sometimes also referred to as "The Big Five" (De Raad, 2000) because "the five factor model forms the basis of one of the most widely used measurement scales, the NEO-Personality Inventory-Revised (NEO-PI-R; Costa & McCrae, 1992)" (Matthews et al., 2009, p. 25). The NEO-PI-R is a comprehensive test that measures a person's score on the five domains and the underlying lower-level traits (see table 2). The NEO-Five Factor Inventory (NEO-FFI), introduced by the same authors, only measures a person’s score on the five broad domains. We administered the NEO-FFI in our study for time related reasons which we will explain in the methods section.

Costa and McCrae and others spent considerable time in researching and integrating the five factors with many other personality schemes (O’ Connor, 2002). Research by Wiggens & Trapnell (1996) and Hofstee, De Raad, & Goldberg (1992) showed that there are clear connections between the five factor model and other interpersonal theories such as Leary’s (1957) interpersonal framework. Thus, when we want to see whether there is a distinct relation between personality and the way behavior of others is interpreted when they annotate with the use of Leary’s interpersonal stance theory, using the five domains and the associated trait facets is a defendable choice. De Raad and Perugini (2002) further reinforce this notion when they state that “The Big Five” serves as a reference model because the five factors capture so much of the matter of personality psychology. There are also critics (see Pervin, 1994; Mischel & Shoda, 1995) and dissenting views (see Block, 1995) on the five factor model, but these are mostly related to questions about the interpretation of the scores on the traits. We intend to look at the interpretations briefly now and into more depth in our result section.

There has been a lot of research describing the five domains and the characteristics or tendencies with which they are associated. We summarized the findings in table 3 and 4 (p. 13) and divided them according to their relation with high and low scores on the NEO-FFI. These tables make discussing each domain and the information about interpreting behavior of others clearer. A few important notions have to be taken in mind so that each domain and their presumed effect on the interpretation of the behavior of others can be understood correctly. Whenever someone is said to score high (or low) on a scale, for instance on conscientiousness, it means that the chance of conscientiousness behavior is higher (or lower) than average (Hoekstra et al., 1996). This distinction is important, because it implies information about the probability of behavior generalized across situations, which is in itself a somewhat problematic notion and forms the basis of a long-lasting dispute about the meaning and interpretation of personality tests. Mischel (1973) and Mischel & Shoda (1995) have pointed to the fact that “*high conscientiousness* people can be consistently orderly in one situation (for instance when working with numbers), or ambitious at work, but not in sports and games” (Hoekstra et al., 1996, p. 26). This cautions us whenever statements are made about people on the basis of the score on a personality test solely because behavior is formed by personality and situations. Researchers agree that it depends on the appraisal of the situation by the person (Hoekstra et al, 1996). In an attempt to answer the question what can be attributed to someone on the basis scores on a personality test we could consider describing personality as “a collection of characteristics of the manner in which situations are distinguished, interpreted and appreciated by the person” (p. 26).

1.4.1. Neuroticism

With these notions in mind we turn to the first domain, Neuroticism. High N-scorers are often associated with negative affect and a proneness to experience fear (see table 3). Furthermore it is found that these people worry often about many things and do not cope with stressful situations very well. Hilbig (2008) found an interesting relation between fast and frugal decision making and neuroticism. Fast and frugal decision making claims that people base inferences on recognition only. Giluk (2009) reinforces this by finding a negative correlation between mindfulness and neuroticism. Mindfulness correlates with conscientiousness, which is related to disciplined, thoughtful and controlled behavior. Germeijs and Verschueren (2011) link indecisiveness to high scores of neuroticism. Although

Table 3. High score characteristics of the five factor model domains

Domain	High score characteristics
Neuroticism	Prone to fear, anxiety, negative feelings like anger, rage frustration, feeling down, embarrassed, feeling guilty, worry often, more unhappy, feeling unsafe, cope less well with stressful situations, indecisiveness
Extraversion	Sociable, like being around others, outgoing, active, good mood, optimistic, like excitement, cheerful, energetic
Openness	Open to new experiences, imagination, intellectual curiosity, sensitivity to aesthetics, eye for own emotional world, independent judgments, playful, flexible, like new unconventional ideas
Agreeableness	Prosocial, helpful, modest, kind, like working with others, relation is often experienced from the other, altruistic, trustful
Conscientiousness	Do what has to be done, reliable, disciplined, controlled, thoughtful, ambitious, orderly, systematic, goal-orientated

Note. Characteristics as described by: Eysenck & Eysenck (1975); Costa & McCrae (1987; 1992; 1997); Hoekstra et al. (1996); Graziano & Eisenberg (1997); Hogan & Ones (1998); Matthews et al. (2009); Germeijs & Verschueren (2011).

Table 4. Low score characteristics of the five factor model domains

Domain	Low score characteristics
Neuroticism	Calm, emotional stable, not easily disturbed, relaxed, approach stressful situations with ease and no tension, negative feelings do not bother them long
Extraversion	Also called introvert; absence of extraversion, not opposite, more independent, inward
Openness	Conventional, stick to things they know consciously, do not look further than is necessary to achieve a goal
Agreeableness	Antagonists, egocentric, look for argument and confrontation, show rejection towards others more easily, competitive, prone to rejection like feelings
Conscientiousness	Less strict with pursuing values and norms, nonchalant, relaxed

Note. Characteristics as described by: Eysenck & Eysenck (1975); Costa & McCrae (1987; 1992; 1997); Hoekstra et al. (1996); Graziano & Eisenberg (1997); Hogan & Ones (1998); Matthews et al. (2009); Germeijs & Verschueren (2011).

our study did not cause any stress or direct fear, we might argue that the association with an, in video shown, negative affect could lead high N-scorers to interpret others behavior more opposed than together (see figure 1 p. 9) in comparison to the other domains, whereas the relation with indecisiveness and fast and frugal decision making may cause the interpretations to be scattered and random. Low N-scorers on the other hand tend to be emotional stable and self-assured (Tellegen, 1985; table 4). Their relaxed behavior may not directly link to distinct preferences toward interpretations. We would however expect a clear difference between high and low N-scorers and the way they interpret behavior.

1.4.2. Extraversion

Studies by Eysenck & Eysenck (1975), Tellegen (1985) and Costa and McCrae (1987) show that people who score high on Extraversion are sociable, optimistic, assertive, energetic and socially dominant. They enjoy being around others and seek harmony. When a conflict is imminent high E-scorers will attempt to sooth the other party and try to downplay the argument (Rosenthal, 1983). They are more likely to engage in conversations (Thorne, 1987; Argyle, Martin, & Crosland, 1989) and are aware of their surroundings. So it might be expected that they will interpret behaviors of others more connected (together) rather than opposed and possibly view the detective more often as leading or helping (see figure 1 and 2, p. 9), whilst interpreting the behavior of the suspect more cooperative and depended. Lucas and Diener (2000) argue that the higher reward sensitivity of high E-scorers makes it more likely that they will seek social situations because they are primarily rewarding. Introversion (Jung, 1923), or scoring low on this domain, has to be regarded as the absence of extraversion, not the opposite (Costa & McCrae, 1987). Low E-scorers are not necessarily uninviting but more reserved, distant and independent. Their focus is inward to their own feelings, actions, and thoughts (Hoekstra et al., 1996). What this means in relation to interpreting behaviors of others is more difficult. Eysenck and Eysenck (1985) suggest that when introverts are highly neurotic that they are more prone to emotional disturbances. As already mentioned, high N-scorers suffer more from indecisiveness and fast and frugal decision making based on recognition only which could show as more scattered ratings.

1.4.3. Openness

The full name that McCrae & Costa (1997) ascribe to the next domain, Openness, is “openness to new experiences, which has to be interpreted more like an attitude” (Hoekstra et al., 1996, p. 32) High O-scorers pay conscious attention and are non-judgmental (Kabat-Zinn, 1994). They live in the present moment, accept values that might be different to their own and have richer experiences than low O-scorers. When interacting with others they are understanding, adapting and favor an egalitarian approach (McCrae, 1996). This might lead them to interpret the behavior of the suspect in our study more often dominant (e.g. in pursue of their own interests) than submissive (adhering to the detective's questions). Low O-scorers on the other hand are more conventional. They stick to what they know and do not look further than is necessary to achieve a goal (Hoekstra et al., 1996). Understanding how high O-scorers interpret behavior of others' is not as easy as it seems. We know that they are more willing to look at the behavior from the others point of view (as do people who score high on Agreeableness) but deducing how they interpret that behavior might considerably change between high O-scorers.

1.4.4. Agreeableness

People that score high on Agreeableness are prosocial. They are very concerned with the perspective of the other because to them that is what matters (Graziano & Eisenberg, 1997). High A-scorers are helpful, kind and prefer to work with others (Wood & Bell, 2008). We expect that they will interpret the behavior of others more frequently as together, rather than opposed and more centered along the horizontal axis (e.g. helpful and cooperative) rather than the vertical axis (e.g. leading and depended, see figure 1, p. 9) (Smets, 2011). Because the mindset of a high A-scorer is on the others' perspective they may be more inclined to interpret behavior as a cry for help, even if that behavior is considered more opposed by the majority. "Low A-scorers are antagonistic and egocentric; they look for the argument and confrontation with others and show their rejection or aggression toward others easier. Their attitude is more competitive than cooperative" (Hoekstra et al., 1996, p. 33). They might be more inclined to interpret the behavior of others as opposed, however this interrelationship is not strongly supported (Jones & Melcher, 1982).

1.4.5. Conscientiousness

The last domain, Conscientiousness, is directly related to conscious and deliberate behavior (Hoekstra et al, 1996). High C-scorers are disciplined, thoughtful and controlled. They do what has to be done because of their goal-orientated nature. They perform better than average in interpersonal functioning because they are less prone to victimization (Jensen-Campbell & Malcom, 2007). They can plan, prioritize and delay gratification in order to achieve what they want. Low C-scorers on the other hand are more carefree and less bothered with punctuality. Consequences are not considered and they have more difficulty to inhibit impulses (Costa & McCrea, 1992, Lee & Ashton, 2006). Theory is not clear on how this relates to interpreting behavior of others and is subject to further research, to which we aim to contribute.

1.5. Hypotheses

The need of human annotation of video content as means of validating interpersonal theories in the development of an affective conversational model for a virtual character has already been explained in several studies (see Vaassen & Daelemans, 2010; 2011; Bruijnes, 2013; op den Akker et al., 2013). The current research aims to contribute to that development by looking deeper in the differences between people in an annotation task, because as op den Akker et al. (2013) mentioned, there are a lot of fuzzy notions when people, who are subjective by nature, have to annotate video material with fixed labels as was the case with Leary's Rose. Although op den Akker et al. provided extensive information how to use and interpret the eight interpersonal behaviors (see op den Akker et al., 2013, p. 197) there still remained considerable disagreement of the interpretation of the utterances. The current study aims to explain the differences by looking at the relation of the personality of the annotators and the choices they make when annotating video material. Although there has been considerable research on the relation between Leary's Rose and other personality theories like the five factor model (see Hofstee, De Raad, & Goldberg, 1992; Wiggens & Trapnell, 1996; and O' Connor, 2002), we are not aware of studies in which differences in annotating video material with the use of Leary's Rose was explained through participants' scores on a personality test. Hypothesis one and two are therefore partly based on personality theories we described and partly a product of our observations and ideas about the subject. Hypothesis three is based solely on personality theories we described earlier.

We looked at the differences between a group of detectives and non-detectives and their answers on an annotation task. Both groups also completed a personality test to see if there were significant differences in the scores on five broad personality domains. Finally we wanted to see if existing theories about five personality domains could be used to explain specific relations with dimensions of Leary's Rose (see pp. 10 -15 of this paper).

We are first and foremost interested in the relation between personality and how behavior of others is interpreted when detectives and non-detectives label that behavior in police interrogations.

Our first hypothesis is that:
There is a difference how detectives and non-detectives annotate behavior of the policeman and interviewee in police interrogations.

Our second hypothesis is that:
There is a difference in the scores on the NEO-FFI between detectives and non-detectives.

Our third hypothesis is based on the personality domains and the theories we described:
There is a positive or negative relation between high or low scores on each personality domain and a specific quadrant or octant of Leary's Rose, more specifically:

a) *There is a positive relation between N-scores and behavior that is labeled as opposed.*
b) *There is a positive relation between E-scores and behavior that is labeled as "leading" and "helping" (see figure 1 p. 9).*
c) *There is a positive relation between A-scores and behavior that is labeled as "helping" and "cooperative" (see figure 1 p. 9)*

2. Methods

2.1. Participants

Two groups participated in this study, 17 detectives that were enrolled in a course at the Police Academy and 17 non-detectives; of which 8 were students from the University of Twente. The total number of participants, 34, was not decided beforehand but limited to the number of detectives in the course at the Police Academy. The detectives were aged between 26 and 57 years (*M* = 40.50, *SD* = 9.07), 9 male and 3 female. 7 Detectives did not provide their age, 5 did not provide their gender. Furthermore, the years of service ranged between 5 and 37 (*M* = 16.10, *SD* = 11.77), 7 detectives did not provide their years of service. The non-detectives were aged between 21 and 55 years (*M* = 31.35, *SD* = 11.68), 10 male and 7 female. There were no financial rewards for participating in the study, although 1 student received a credit[4].

We also checked whether or not the participants had experience with annotating video content and personality tests. One detective had prior experience with annotation, 9 did not and 7 did not answer the question. Of the non-detectives 6 had prior experience with

[4] At the University of Twente students have to acquire a certain amount of credits to complete their Bachelor's degree. These credits are rewarded by participating in studies.

annotation, of which 1 with police interrogations as in our study, 11 had no prior experience. 6 detectives had prior experience with personality tests, 4 did not and 7 refrained from answering. Fourteen non-detectives had prior experience with personality tests, 3 did not.

We were bound to the group detectives that were enlisted in the course at the time of our study which prohibited random selection. These detectives had enlisted in the course voluntary, although it was mentioned that the Police Academy encouraged taking this course to raise the interrogations skills. The group was demographically diverse with regard to age, gender and years of service. Non-detectives were chosen as follows: an email was sent to a large group of students from the University of Twente explaining the interview and inviting participation. Furthermore the study was posted on SONA-systems, the designated website of the university to invite participants. Because we wanted to match the number of participants from the detectives group, we also randomly asked people outside the university to participate. The group non-detectives were also demographically diverse in regard to age and gender. All participants were given an informed consent form before the start of the study that stated clearly that their participation was voluntarily and that they could withdraw from the study at any time.

2.2. Pilot

In order to check whether the instructions in our study were easy to understand we presented them to 5 people outside the University and Police Academy. These people were not asked to complete the NEO-FFI or to annotate the video material, but only to read the instructions and annotate one example video. This video was not used in our study. Their input was discussed and some moderations were made to the instructions. See appendix 1 to 6 for the instructions used in our study.

2.3. Materials

The Police Academy granted us cooperation, but we were asked to honor a certain time limit. We therefor administered the NEO-FFI, which is the short version of the NEO-PI-R and only measures the five broad personality domains: *Neuroticism*, *Extraversion*, *Openness*, *Agreeableness* and *Conscientiousness* (Costa & McCrae, 1992). The Dutch/Flemish adaption of this inventory employed (Hoekstra et al., 1996). The NEO-FFI consists of 60 items which take approximately 20 minutes to complete. Participants rate how well the statements represent their opinion on a 1 – 5 Likert scale (from totally disagree to completely agree).

Leary’s (1957) framework of interpersonal stance was used to annotate the behavior of the interrogator and interviewee in the video clips. Participants received a verbal explanation of Leary’s Rose alongside an informational folder and rating form. The video material was restricted to two interrogations that had already been shown on national television a few years earlier and were therefor publically available. Both interrogations coincidentally concerned murder cases. A third interrogation was used with permission from the Police Academy, this was not a real case, but a recorded role play between a detective and actor. The interrogations where cut in 42 fragments of which 20 were chosen randomly, numbered 1 to 20. The video clips consisted of 8 to 31 second fragments of a moment in the interview, containing sufficient material to annotate on. Each video clip faded in and out. Two different video clips served as trials: one showed exaggerated examples of specific stances from Leary’s Rose, the other showed a fragment of a television show interview. Two more informational videos about serious gaming showed an ECA being used in a military training program. Local available computers and headsets were used and we brought one laptop as a back-up. 11 USB-sticks (2 reserve) contained the video clips. See appendix 1 to 6 for the materials that were handed out.

2.4. Procedure

To a large extend the annotation task followed the procedure from op den Akker et al. (2013). The current study differed in that it pre-segmented individual and separated fragments that could be annotated manually. This design was chosen to accommodate the available time and make the annotation task easier to do and understand. This was a specific request from the Police Academy. The study consisted of two parts: a personality test and an annotation task. To avoid a possible bias due to completing either part first, both groups were divided and started either with the personality test or the annotation task. All data was collected anonymously, only noting to which group a participant belonged. Relating the test results back to a participant was not part of our research design, nor did we need it to confirm or denounce our hypothesis. To a lesser extend it helped to gain cooperation of the participants and reassure to them that there were no personal consequences attached to the personality test.

Together with the lecturers from the Police Academy we discussed and agreed that the experiment could best be administered when they treated Leary’s Rose in class. This ensured that the teacher could prepare and give her lecture as she best saw fit and so that we did not interfere with the course and workload of the detectives too much. At the start of the class she shortly introduced the researcher, who then explained the intention of administering a test after the lecture on Leary's Rose. The detectives had not previously met the researcher (we were an auditing guest in a different class, but the same course). The researcher took precise notes how the lecturer introduced and lectured the class on Leary's Rose. A similar verbal introduction was later given to the non-detectives group. This was anticipated in advance and ensured that the test was introduced and administered under similar circumstances in the non-detectives group.

After finishing her lecture the researcher was allowed in front of the class. He briefly highlighted some details on Leary's theory, strictly following the informational bundle that was handed out. The bundle contained all the instructions and information they needed to complete the experiment (see appendix 1 to 6). An introduction was provided to the detectives about the master Human Factors and Engineering and how it related to the current study. Then the informed consent form was explained, followed by an outline of the experiment. An additional explanation about the NEO-FFI was given because assumptions with which a participant completes the test influence the way in which the scores can be interpreted (Hoekstra et al., 1996). To assure correct interpretation participants were remembered that they participated in a study that was done to aid the development of a digital tutoring system that could be used by the police and that the results of the NEO-FFI might support that development. We explained that there were no personal consequences attached to the test result, therefore there was no possibility of providing individual feedback. The detectives were instructed to read the standardized instruction to the personality test before starting (see Hoekstra et al., 1996).

The group that started with the personality test remained in the classroom with the teacher while the other group went to a computer room accompanied by the researcher. Each detective received a headset and was assigned to a computer on which the folder containing the video clips was already opened. The group was explained that they would annotate 20 video clips of a police interrogation with the use of Leary's Rose. The informational folder contained exact instructions on the annotation task (see appendix 1 to 6). The group was told to read the instructions, complete the example trials and begin if they had no further questions. Upon completing the annotation task, the detectives were asked to refrain from discussing the test with the group that had completed the personality test. The groups then changed rooms and completed the other test.

We reserved a computer room at the university with the intention to repeat the exact same procedure. However, eight students (non-detectives) turned up separately from each other which lead us to slightly alter our approach. We followed the previously described procedure, but did so per participant. The remaining tests were conducted in a similar manner elsewhere on the university’s campus and in a few cases at a participants’ home (under aforementioned controlled circumstances). Of all non-detective participants, half took the personality test first followed by the annotation task and vice versa.

3. Results

3.1. Descriptive statistics

Descriptive statistics for scores on each of the five domains from the NEO-FFI are displayed in table 5. Each participant answered all the questions. We found no instances of faking, random answering or acquiescence. The raw scores on the personality test were transformed to norm scores by *gender* in a *research context* (Hoekstra et al., 1996). We chose gender due to a relatively large number of instances where no age was recorded. There were five cases in which gender data was also missing (all cases related to detectives). We transformed the raw scores from those instances with the norm-table for both genders. We first plotted boxplots to get a general idea about the distribution between both groups. Only Agreeableness seemed to differ between the groups (see figure 3 p. 22).

Each participant could allocate more than 1 stance to the interrogator and interviewee. After inspecting the data we decided that we could not interpret with certainty what it meant when a participant allocated more than 1 stance to either the interrogator or interviewee. We therefor decided to only look at the stance that each participant labeled first. We will discuss this under limitations. Table 7 (p.23) and 8 (p.24) display the total number of labeled stances for interviewee and interrogator respectively. Preliminary boxplots showed that both groups only differed in the allocation of the defiant stance to the interviewee (see figure 4 p. 23).

Table 5. Descriptive statistics NEO-FFI norm-scores by gender in research context

Variable	*M*	SD	Min	Max	*N*
Neuroticism	4.18 (*4.59*)	2.13 (*1.81*)	1 (*1*)	8 (*7*)	17 (*17*)
Extraversion	6.00 (*6.41*)	2.26 (*1.91*)	1 (*3*)	9 (*9*)	17 (*17*)
Openness	5.76 (*6.12*)	1.82 (*2.12*)	2 (*2*)	9 (*9*)	17 (*17*)
Agreeableness	5.82 (*3.94*)	1.91 (*1.98*)	2 (*1*)	8 (*8*)	17 (*17*)
Conscientiousness	5.35 (*4.82*)	2.00 (*2.56*)	2 (*1*)	9 (*9*)	17 (*17*)

Note. Detectives and *Non-detectives*

Table 6. How to interpret the scores on the personality domains

Score	Interpretation
1	Extremely low
2 & 3	Low
4	Low average (*laag gemiddeld*)
5	Average
6	High average (*hoog geniddeld*)
7 & 8	High
9	Extremely high

3.2.Hypotheses 1, 2 and 3

Our research question and primary hypothesis concerned the difference on the annotation task between the detectives and non-detectives. Table 7 (p. 23) and 8 (p. 24) display noticeable differences between ratings of detectives and non-detectives on *defiant* and *competitive* behavior for the interviewee and *leading*, *cooperative* and *defiant* behavior for the interrogator. We deleted the *neutral* stance from our results because it was only chosen in 2 (out of possible 1360) instances. From the boxplots we learned that only the mean ratings on *defiant* behavior of the interviewee seemed to differ greatly between both groups. Independent samples *t* test were performed comparing the mean scores of the annotation task between detectives and non-detectives.

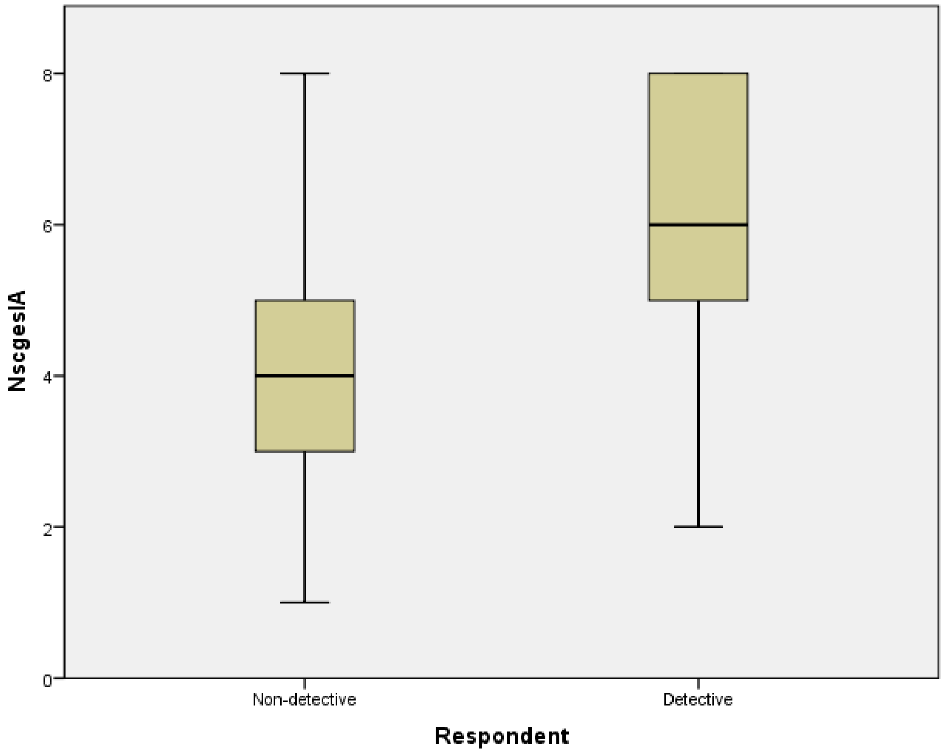

Figure 3. Mean differences between non-detectives (left) and detectives (right) on scale Agreeableness.

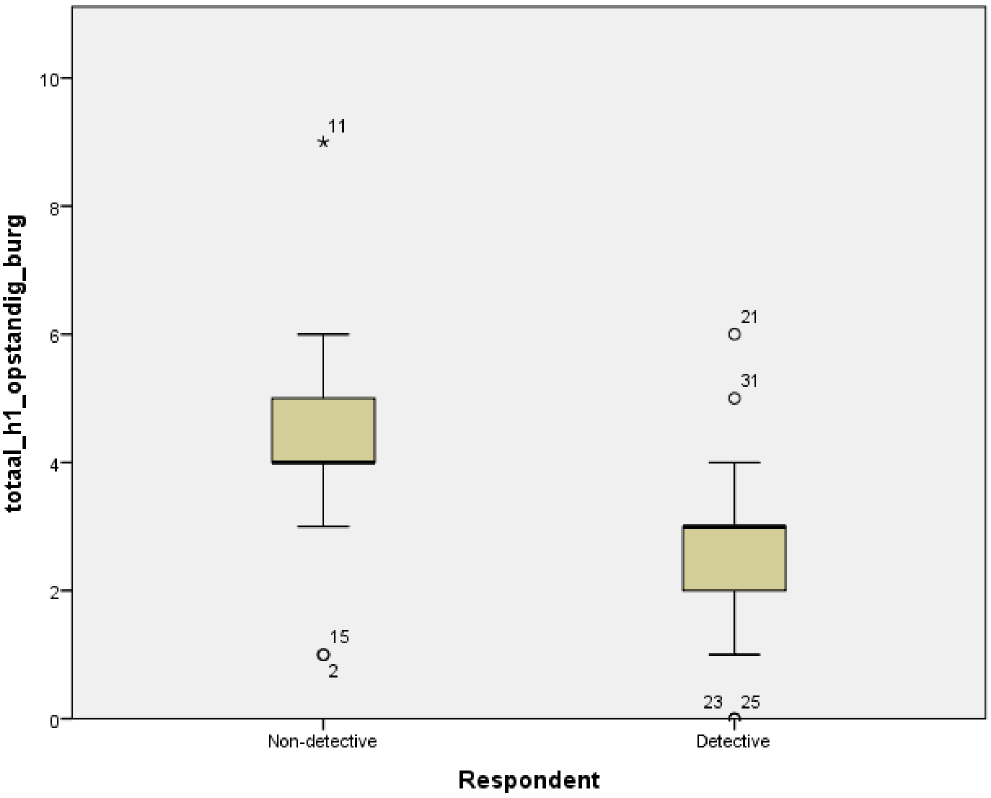

Figure 4. Mean differences between non-detectives (left) and detectives (right) on the allocation of a defiant stance from the interviewee.

Table 7. Total stance allocating for interviewee across fragments

Variable	Detectives	*Non-detectives*
Leading	22	*16*
Helping	14	*11*
Cooperative	115	*110*
Depend	46	*40*
Withdrawn	55	*54*
Defiant	47	*73*
Aggression	22	*16*
Compete	8	*17*

Note. Only the first allocated stance in each video clip was used in our analyses.

We compared the total allocated first annotated behaviors from the interviewee and interrogator. We did not compare every single video clip because they were not a controlled factor, but rather chosen randomly. Interpreting any possible differences in the ratings on each video clip was deemed obsolete. Non-detectives (*M* = 4.29, *SD* = 1.83, *N* = 17) rated defiant behavior of the interviewee significantly higher than detectives (*M* = 2.77, *SD* = 1.56, *N* = 17), *t*(32) = 2.62, *p* = .013. We also looked at the other mean differences for *competitive*

Table 8. Total stance allocating for interrogator across fragments

Variable	Detectives	*Non-detectives*
Leading	150	*130*
Helping	72	*63*
Cooperative	34	*49*
Depend	14	*12*
Withdrawn	1	*3*
Defiant	2	*11*
Aggression	41	*46*
Compete	17	*20*

Note. Only the first allocated stance in each video clip was used in our analyses.

behavior for the interviewee and *leading*, *cooperative* and *defiant* behavior for the interrogator, these all were non-significant.

Our second hypothesis concerned the possible differences on the scores of the personality domains between the groups. Independent samples *t* test were performed comparing the mean scores of detectives and non-detectives on the five domains of personality. As we predicted from the boxplot (figure 3) the scores on Agreeableness differed significantly. Non-detectives (*M* = 3.94, *SD* = 1.98, *N* = 17) scored significantly lower on Agreeableness than detectives (*M* = 5.82, *SD* = 1.91, *N* = 17), *t*(32) = -2.82, *p* = .008. Both groups did not differ significantly from each other on the other domains, *t*(32) = .61, *p* = .55 for Neuroticism, *t*(32) = .57, *p* = .57 for Extraversion, *t*(32) = .52, *p* = .61 for Openness and *t*(32) = -.67, *p* = .51 for Conscientiousness (for associated *M* and *SD* see table 6 p. 21). Our hypothesis that there are differences between the scores on the NEO-FFI domains in both groups (*hypothesis 2*) is only partially supported. The mean score of the detectives on Agreeableness was 5.82 (see table 6). According to Hoekstra et al. (1996) this score can be considered average. Detrick and Chibnall (2013) found similar average scores among 288 police officer applicants. Although Detrick and Chibnall noted that an average score on Openness and Agreeableness is favorable, they concluded that police officers should preferably have low scores on Neuroticism and high scores on Extraversion and Conscientiousness. Our group detectives had high average scores on Extraversion (see table 6 p. 22 for individual differences) and average scores on Openness and Agreeableness (see table 6 p. 22). Another study on differences between police officers and a reference group found significant differences between Agreeableness, Conscientiousness and Neuroticism and Openness, but not Extraversion (Abrahamsen & Strype, 2010). In their study police officers also scored higher on Agreeableness.

Before continuing to the third hypothesis we would like to make a minor sidestep. Hypothesis one showed that non-detectives rated defiant behavior of the interviewee significantly higher than detectives. The second hypothesis showed that non-detectives scored significantly lower than detectives on Agreeableness. Our initial research question is concerned with the differences on the annotation task between both groups and the differences between the scores on the personality test. To test whether there is an interaction effect between the scores of defiant behavior and the scores of Agreeableness we performed a Generalized Linear Models (GLM) analysis with a linear model. Schmettow (2013) introduced GLM analysis during a lecture on research methods. "The generalized linear model expands the general linear model so that the dependent variable is linearly related to the factors and covariates via a specified link function" (IBM, n.d. p. 46). Furthermore the model can handle a non-normal distribution of the dependent variable.

We first performed a GLM analysis with a log linked Poisson distribution with the counts of *defiant* behavior from the interviewee as dependent variable and expected to find results similar to our *t* test in hypothesis one. Respondent was used as categorical predictor in the model. A GLM analysis with a log linked Poisson distribution allows the dependent variable to be discrete and counted. To support the use of a Poisson distribution we looked at the residual distribution of the number of times that behavior of the interviewee was rated *defiant* (see histogram, figure 5, p. 26). Results show a positive skewness value of .241 (SE = .403) and a positive Kurtosis value of 1.363 (SE = .788). As already mentioned the use of a Poisson distribution is allows for a non-normal distribution.

Table 9 shows parameter estimates for our GLM. When using this type of model (Poisson distributed model with a log link) the regressions coefficients (β) are log transformed. In order to make sense of the values we have to exponentiate them. The intercept in table 9 is the expected number of ratings of *defiant* behavior of the interviewee. From our analysis we see that there is a significant positive regression coefficient for non-detectives ($\chi 2$=5.543, p = .019). So non-detectives rate the behavior of the interviewee, exp(1.017) – (exp(1.017)* exp(.440)) = 1.55 times more *defiant* than the detectives. As expected this result is in line with the *t* test we performed.

We introduced the GLM model because we wanted to see whether the score on Agreeableness has an effect on the ratings of *defiant* behavior of the interviewee. This appears not to be the case. We performed a GLM analysis with the counts of *defiant* behavior from the interviewee as dependent variable, respondent as categorical predictor and the scores on

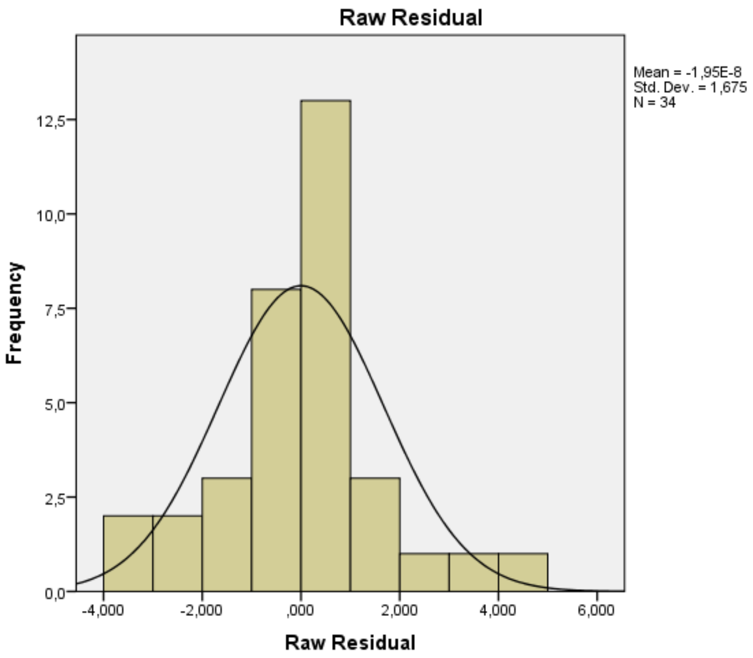

Figure 5. Histogram of the residual distribution of the linear model.

Table 9. Parameter estimates in the GLM analysis with the detective group as reference

		95% CI			
Parmeter	β (SE)	Lower	Upper	χ2	p
Intercept	1.017 (.145)	.731	1.303	48.605	<.0005
Respondent					
Non-detective	.440 (.187)	.074	.807	5.543	= .019
Detective	0				

Note. β-values are log transformed. Dependent variable *defiant* behavior of the interviewee.

Agreeableness as covariate predictor. There results were non-significant, see table 10 (non-detectives, $\chi 2$= .382, *p* = .066, norm score Agreeableness, $\chi 2$= -.031, *p* = .519).

Table 10. Parameter estimates in the GLM analysis with the detective group as reference

		95% CI			
Parmeter	β (SE)	Lower	Upper	χ2	p
Intercept	1.196 (.311)	.586	1.806	14.773	<.0005
Respondent					
Non-detective	.382 (.208)	-.026	.789	3.372	= .066
Detective	0				
NscA	-.031 (.048)	-.125	.063	.0416	= .519

Note. β-values are log transformed. Dependent variable *defiant* behavior of the interviewee.
NscA: Norm score by gender on Agreeableness.

We also tested if there was an interaction effect between the scores on Agreeableness of both groups and the ratings of *defiant* behavior of the interviewee. We compared the goodness of fit of the proposed model by looking at the Akaike's Information Criterion (AIC) with and without an interaction effect. In the model without the interaction effect the AIC score was 137.08. We ran a second analysis with the interaction model; this showed an AIC of 138.18. A lower AIC score indicates a better goodness of fit of the model (Schmettow, 2013). Judging from the AIC solely we could argue against adding an interaction effect to the model. Because the AIC differed only slightly we chose to include an interaction effect and test the model. We performed a GLM analysis with the counts of *defiant* behavior from the interviewee as dependent variable, respondent as categorical predictor and the scores on Agreeableness as covariate predictor. The intercept in table 11 is the expected number of ratings of *defiant* behaviors of the interviewee. Interaction represents a moderation effect which means that the effect of one variable depends on the level of another variable. The results were also non-significant (see table 11). We conclude that our first hypothesis is only slightly supported. We only found a significant difference between the detectives and non-detectives and the ratings of *defiant* behavior of the interviewee.

Table 11. Parameter estimates in the GLM analysis with the detective group as reference

		95% CI			
Parmeter	β (SE)	Lower	Upper	χ2	p
Intercept	1.520 (.448)	.641	2.399	11.489	= .001
Respondent					
Non-detective	-.082 (.522)	-1.106	.942	.025	= .875
Detective	0				
NscA	-.89 (.077)	-.240	.062	1.331	= .249
Non-det*NscA	.094 (.098)	-.099	.286	.911	= .340
Det*NsA	0				

Note. β-values are log transformed. Dependent variable *defiant* behavior of the interviewee.
NscA: Norm score by gender on Agreeableness.

We conducted numerous GLM analyses between ratings of both groups without finding significant differences. We decided not to mention them here for readability purposes. We will turn back to this in our conclusions and limitations.

Our third hypothesis was chosen to see if the results on the annotation task related to the interpersonal theories of behavior we described. We conducted correlation analysis on the relation between Neuroticism and behavior that was annotated as *opposed* (hypothesis 3a), Extraversion and behavior that was annotated as being *leading* and *helping* (hypothesis 3b) and Agreeableness and behavior that was annotated as being *helping* and *cooperative* (hypothesis 3c). The correlations are shown in table 12. We did not look at differences between the groups as previously mentioned. No relations were found to support our hypotheses. We will draw conclusions in the next section.

Table 12. Correlations between personality and annotated behavior.

	Opposed		Leading and helping		Helping and cooperative	
Variable	1.	2.	1.	2.	1.	2.
Neuroticism	-.031 (.863)	.293 (.093)				
Extraversion			-.099 (.576)	-.002 (.991)		
Agreeableness					.050 (.799)	-.052 (.772)

Note. Pearson correlation and (significance). 1 = interviewee, 2 = interrogator.
Blank spots were not analyzed.

Our research question and three initial hypotheses are only slightly supported and leave room for many different interpretations. We will focus on explaining what we can take away from our study. But before we do so would shortly like to look at our annotation task differently. In the op den Akker et al. (2013) study notions were made about the agreement in the group annotators. Since our results indicated that there were not as many differences between the groups as we thought, we aimed to find whether they answered in a similar fashion. In addition to our research and aiming to contribute to the development of a system of stance recognition for the ECA we want to pay attention to the annotation task.

Op den Akker et al. (2004) measured inter-annotator agreement using Krippendorff's alpha. This is a "general method for comparing arbitrary number of annotators allowing different distance metrics on the label set (Krippendorff 2004)" (p. 200). They computed a Boolean metric alpha of 0.24 for all annotators. A Boolean metric means that two annotators label behavior of the interrogator or interviewee in a video clip equal (distance is 0) or not (distance is 1). In other words it shows the strength of the overall inter-annotator agreement. Our results are approximately equal, Krippendorff's alpha in the group detectives was $\alpha = .22$ ($N = 17$) and $\alpha = .20$ ($N = 17$) in the group non-detectives. Although minor, it appears that there is more inter-annotator agreement in the group detectives compared to non-detectives, but less compared to the group annotators in the op den Akker et al. study.

We also calculated a distant measure (KL), ranging from 0 and (2 ln 2) = 1.386 based on the Kullback–Leibler divergence (see Kullback & Leibler, 1951; Kullback, 1997) to attain more information about the differences between the annotations of the interviewee and interrogator in the video clips. KL was calculated between two smoothed probability distributions. Smoothing avoids null-values in the probability distribution. A simple add K smoothing was used with K = 0.1, also known as a Laplacian smoothing (R. op den Akker, personal communication, June 25, 2014). Four distance measures were calculated (see table 13 and 14 p. 27) for both groups. The measures tell us something about how much the probability distribution deviates from random (uniform distribution) in each group. Ideally the measure in the first three columns will be further from 0 than the last, since it shows the difference between the groups. This appears to be the case. Means and standard deviations are also presented in table 15, p. 31.

Interpreting these numbers might be cumbersome, but we generally see that the numbers in the first three columns differ from random quite profoundly. Put in other words, the KL divergence measures in table 13 and 14 indicate that the annotators do no allocate the stances randomly to the interviewee and interrogator, but rather that there is some agreement between the annotators. From observing the data we see that although there can be quite some differences in allocating stances to the behavior of the interviewee and interrogator, there is also a lot of agreement. Furthermore, there are instances in which the annotators agree on the quadrant (e.g. AT+TA, see figure 1 p. 9) but not on the octants (e.g. *leading or helping*). These agreements show that annotators have a general conception of the sort of behaviors that the interviewee and interrogator display.

We represented four heat maps to get another picture of the data. Two heat maps were composed for the ratings on the behavior of the interviewee; one with the smallest KL and one with the largest KL. The same was done for the behavior of the interrogator. We chose to look at the fourth column, but one has to take care to interpret the data correctly. This column represents the distance between the both groups. A low number actually means higher agreement and vice versa. The heat map shows nicely how both ratings compare or differ (see figure 6, 7, 8 and 9, p. 32).

If we take a closer look at table 15 (p. 31) we can see that the mean and standard deviation do not seem to differ greatly from each other. In order to see whether there was a difference we conducted a paired samples *t* test. We interpreted the paired samples *t* test as follows: we have two groups, detectives and non-detectives who annotated stance from the interviewee and interrogator on each video clip. We contrasted each annotation against a random annotation based on a KL divergence distance measure. The correlations then, says something about how a judgment on moment 1 is linked with moment 2. In our case that is a somewhat odd notion, but we argue that because in our case the annotation from detective on moment 1 is linked against the same random as the annotation from the non-detective on moment 1 we can use the paired samples *t* test to see whether there is a difference in the coherence between

Table 13. Distance measures (KL) for ratings on behavior of interviewee

Video	KL (NonD, random)	KL (D, random)	KL (NonD+D, Random)	KL (NonD, D)
1	0,363	0,493	0,399	0,260
2	0,534	0,232	0,344	0,207
3	0,506	0,347	0,417	0,116
4	0,466	0,320	0,365	0,128
5	0,574	0,439	0,489	0,205
6	0,311	0,567	0,377	0,288
7	0,494	0,522	0,455	**0,302**
8	0,366	0,648	0,487	0,158
9	0,458	0,413	0,463	0,041
10	0,221	0,379	0,263	0,188
11	0,491	0,476	0,516	**0,019**
12	0,336	0,365	0,343	0,100
13	0,261	0,145	0,181	0,128
14	0,488	0,506	0,507	0,060
15	0,328	0,466	0,362	0,164
16	0,343	0,510	0,418	0,110
17	0,291	0,336	0,332	0,057
18	0,213	0,117	0,131	0,167
19	0,153	0,149	0,102	0,193
20	0,302	0,140	0,200	0,113

Note. KL distance measures between 0 and 1.386, ratings of first annotated behavior. Distances are measures between annotations and random. A greater value represents a larger distance between the two probability measures.
NonD = Non-detective, D = detective.

the groups. We first conducted a Pitman Morgan test to check the difference between variances. If the Pearson correlation is statistically significant, the variances are considered to be significantly different from each other and performing a paired samples *t* test might not be favorable. In both conditions (e.g. ratings of interrogator and ratings of interviewee) the Pearson correlation was not statistically significant *r*(18) = .61, *p* = .526 and *r*(18) = .4, *p* = .222 respectively. The results of the paired samples *t* test indicated that there is not a convincingly significant difference in the coherence between the groups when annotating the interrogator. The mean for the distances between the non-detective and random (*M* = .390, *SD* = .119) was not significant lower than the mean between the detective and random (*M* = .445, *SD* = .107), *t*(19) = -2.076, *p* = .052. The results of the paired samples *t* test indicated no significant difference in the coherence between the groups when annotating the interviewee. The mean for the distances between the non-detective and random (*M* = .374, *SD* = .119) was partially significant lower than the mean between the detective and random

Table 14. Distance measures (KL) for ratings on behavior of interrogator

	KL	KL	KL	KL
Video	(NonD, random)	(D, random)	(NonD+D, Random)	(NonD, D)
1	0,428	0,526	0,504	**0,033**
2	0,477	0,342	0,383	0,178
3	0,526	0,426	0,479	0,208
4	0,534	0,534	0,570	0,037
5	0,402	0,549	0,487	0,098
6	0,445	0,526	0,426	**0,242**
7	0,169	0,143	0,118	0,128
8	0,371	0,436	0,421	0,100
9	0,466	0,487	0,491	0,097
10	0,186	0,377	0,239	0,209
11	0,371	0,477	0,440	0,089
12	0,411	0,484	0,456	0,085
13	0,562	0,436	0,487	0,182
14	0,466	0,488	0,491	0,089
15	0,412	0,499	0,48	0,036
16	0,221	0,493	0,319	0,190
17	0,540	0,577	0,584	0,125
18	0,203	0,261	0,181	0,233
19	0,412	0,51	0,471	0,142
20	0,366	0,333	0,363	0,065

Note. KL distance measures between 0 and 1.386, ratings of first annotated behavior. Distances are measures between annotations and random. A greater value represents a larger distance between the two probability measures.
NonD = Non-detective, D = detective.

Table 15. Descriptive statistics for distance measures

Variable	*M*	SD	Min	Max	*N*
Dist_NonD_Ran	.375 (*.398*)	.119 (*.119*)	.153 (*.169*)	.574 (*.562*)	20 (*20*)
Dist_D_Ran	.378 (*.445*)	.155 (*.107*)	.117 (*.143*)	.648 (*.577*)	20 (*20*)
Dist_Tot_Ran	.357 (*.420*)	.125 (*.121*)	.102 (*.118*)	.516 (*.584*)	20 (*20*)
Dist_NonD_D	.150 (*.128*)	.078 (*.066*)	.019 (*.033*)	.302 (*.242*)	20 (*20*)

Note. Dist = distance, NonD = non-detective, D = detective, Ran = Random. Shown are distance measures for interviewee and (*interrogator*).

(M = .378, SD =.155), $t(19)$ = -.103, p = .919. Looking at the data from a distance we can say that the coherence between the groups does not differ much. We conclude that, based on the KL divergence measures, there is a difference between the ratings and random, but there are no significant differences between the groups. This underpins our previous mentioned findings.

Variable	Detectives	Non-detectives
Leading	2	2
Helping	1	2
Cooperative	11	12
Depend	2	1
Withdrawn	0	0
Defiant	0	0
Aggression	0	0
Compete	0	0

Figure 6. Heat map video clip 11.
Behavior interviewee. KL distance = .019.
All behavior concentrated in together half of Leary's Rose. *N* = 16 (17)

Variable	Detectives	Non-detectives
Leading	4	0
Helping	0	0
Cooperative	0	0
Depend	0	0
Withdrawn	0	0
Defiant	2	4
Aggression	10	9
Compete	0	4

Figure 7. Heat map video clip 7.
Behavior interviewee. KL distance = .302.
Difference in opinion is largest.
N = 16 (17)

Variable	Detectives	Non-detectives
Leading	11	9
Helping	3	3
Cooperative	3	3
Depend	0	0
Withdrawn	0	0
Defiant	0	1
Aggression	0	0
Compete	0	0

Figure 8. Heat map video clip 1.
Behavior interrogator. KL distance = .033.
Behavior concentrated in the together half of Leary's Rose. *N* = 17 (16)

Variable	Detectives	Non-detectives
Leading	10	12
Helping	3	0
Cooperative	0	1
Depend	0	0
Withdrawn	0	1
Defiant	0	0
Aggression	2	1
Compete	0	1

Figure 9. Heat map video clip 6.
Behavior interrogator. KL distance = .242.
Behavior largely concentrated, but differences exist. *N* = 15 (16)

4. Discussion

4.1. Hypotheses and conclusions

Our hypothesis which made statements about differences between groups, personality and levels of personality domains were only partially supported. Our foremost predicted difference between the group detectives and non-detectives was not found, expect for one octant of Leary's Rose. Non-detectives perceived the behavior of the interviewee significantly more *defiant* than the detectives. We can argue that this could be due to different perspectives on how someone normally behaves in a certain situation. A detective that has experience with interrogations might be used to *defiant* or non-cooperative behavior and have unconsciously developed an internal filter. We do not want to speculate too much, but during our time at the Police Academy we came to realize just how much hardship these detectives encounter during their careers. They told us that for them, drama and traumatic events are part of the job and they have to find a way to cope with it and downplay the sharp edges. We are by no means saying that they do not care (because they are indeed very involved), but we hint that a suspect behaving defiantly might just be perceived as a common factor. From other fields we know that workers take more risks and ignore safety measures from time to time because they are no longer perceived as risks due to habitation (Wagenaar, 1992).

Stating that detectives are involved and care about others is reinforced when we look at the findings of our second hypothesis, which was partially supported. It showed that detectives score significantly higher than non-detectives on Agreeableness, but furthermore there appeared to be no differences between the two groups. We found that the scores on Agreeableness coincide with those found in the study conducted by Detrick and Chibnall (2013). They state that, ideally, police officers score average on Agreeableness and Openness. In our sample this appeared to be confirmed. Overall our results match the suggested profiles of Detrick and Chibnall (2013), but this is beyond the scope of our study.

When we proposed our initial research question it seemed straightforward. Would two groups that seem to differ on face value and due to their occupations annotate behavior of others differently? As we have seen from our results the difference based on personality and occupation, in this case detective or non-detective, does not lead to a great deal of variance, at least, not when conducting an annotation task. This is not the whole picture though, because there are differences when people annotate behavior and these differences appear in both groups. Omarzu & Harvey (2012) explain that we have great difficulty interpreting behavior of someone else when they show conflicting behaviors, e.g. very aggressive on a personal question in one instance and very cooperative on a similar personal question in another instance. Our study contained materials from three cases, in which two suspects were murder suspects and one was brought in because she supposedly had molested her neighbor. It could be imagined that in these conditions the interviewees showed conflicting behavior. Although each participant watched the same video clips and was therefore exposed to the same possible conflicting behavior, we might argue that some participants handled these conflicting behaviors differently.

Our third hypothesis intended to show how much personality theories about interpersonal behavior could be related to interpreting behavior on an annotation task. In hindsight our predictions may have been somewhat optimistic. In order to reach better conclusions about how others' behavior is interpreted, this study partially missed out on one important feature. Communication, as we know, involves two distinguished dimensions: a verbal one, e.g. the words we use or 'what is said' and a non-verbal one, e.g. the sounds we make, body- and facial expressions, intonation and gestures or 'how it is said'(Carson, 1969). Because we were limited to the materials that are publically available not all video clips showed the faces and/or complete bodies of the suspect/witness and detective and since this emphasizes the focus on what was said, this could have led to different interpretations (Omarzu & Harvey, 2012). This might not be completely true since participants could hear how things were said, we would argue that there might be a difference in the relation to the interpersonal theories.

4.2. Implications and critical notions

In our study we were interested in annotations of suspects and/or witnesses because our goal is to gain more insights in the perceptions of stance recognition. We were therefore bound to video material that was publically available and given to us by the Police Academy. We had no control over the content, quality, and camera angles. We are curious if there is a difference when other material is used. Also, the Police Academy mentioned that there focus is on improving witness interrogations. Ideally we would want to redo a study but then focus on how witnesses are annotated.

One might argue that witnesses are not good material for stance computation because they are more cooperative and therefor little variance might be found. In our opinion this is not necessarily the case because Leary (1957) explained that when people interact with each other all interpersonal behavior is used. In fact we believe that not every witness is as cooperative as one might think. We can think of many reasons why a witness does not really want to cooperate. The suspect may live in the same area, and therefore the suspect might feel in danger or there could be personal face threating issues. Noticing these behaviors can be very important in an interview, especially, when we take into consideration that a deposition from a witness might be used in trial. However, we would also need permission for the use of the video material and this might not be easy.

Another limitation of our study was the number of participants. We relied on the cooperation and possibilities of the Police Academy and detectives in class. They were willing to cooperate but time was limited because we conducted the test during a class. We needed more participants to reliably make sense of the differences between high and low scores on the five domains. In most conditions only very few participants scored high or low. We therefore refrained from conducting extensive analyses between those scores.

We have to note that the way we looked at personality through scores on the five domains is only partly correct and defensible. We explained this thoroughly already, but would like to raise this awareness again. We argued that we were only interested in broad differences and whether there was a connection to the way people interpret behavior of others when performing a annotation task we could justify conducting our study. We did not find any direct relations, but would encourage that if in future studies such relation should be found, they be handled with care. Reality is that most people score about average and will show mixed behavior and it is not correct to judge someone just by their scores (Hoekstra et al, 1996). In case of apparent relations we would have recommended further analysis to find a meaningful interpretation (Mischel, 1973; Mischel & Shoda, 1995).

Our final notion is more of a general kind because each participant was subject to this bias due to our research design. When someone acts as an observer he or she is more focused on what a person says and how he acts. Observers are then more likely to attribute the behavior to the actors' overall personality rather than to ascribe that behavior to the environment in which it occurred. This is also known as the actor-observer bias (Heider, 1958; Ross, 1977). We mention this because it might relate to some of the problem of inter-rater agreement. When there is more focus on the person than the environment different interpretations might easily be formed by different annotators and they might stick across the entire duration of the annotation task.

4.3. Closing statements

We aimed to contribute to the development of a system of stance recognition by looking how personality related to interpreting behavior. Although there are no immediate results, we still feel to have contributed to the field. From our results we might argue that differences in occupations or personality seem to have little or no effect on the performance and attributions in an annotation task. This information could serve well when looking at other groups to annotate material, because the development of the system depends on numerous annotations. We learned a great deal from our time at the Police Academy and feel that the gap between developing and implementing new ways to learn or enrich, in our case, interrogation skills, relies on close cooperation between both the developer and user. We were pleased to find that once we had introduced the detectives to the possibilities of also using an ECA they were very interested and could picture themselves working with it. We are indeed very grateful for their enthusiastic approach and cooperation.

References

Abrahamsen, S., & Strype, J. (2010). Are they all the same? Norwegian police officers' personality characteristics and tactics of conflict resolution. *Policing & Society*, *20*(1), 99-123.

Akker, op den, R., Bruijnes, M., Peters, R., & Krikke, T. (2013). Interpersonal stance in police interviews: content analysis. *Computational Linguistics in the Netherlands Journal (CLIN Journal), 3*, 193-216.

Argyle, M., Martin, M., & Crossland, J. (1989). Happiness as a function of personality and social encounters. *Recent advances in social psychology: An international perspective*, 189-203.

Becker, W. C., & Krug, R. S. (1964). A circumplex model for social behavior in children. *Child Development*, 371-396.

Block, J. (1995). A contrarian view of the five-factor approach to personality description. *Psychological bulletin*, *117*(2), 187.

Borgatta, E. F., Cottrell Jr, L. S., & Mann, J. H. (1958). The spectrum of individual interaction characteristics: An inter-dimensional analysis. *Psychological Reports*, 4(3), 279-319.

Bruijnes, M. (2013, September). Affective conversational models: Interpersonal stance in a police interview context. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on* (pp. 624-629). IEEE.

Carson, R. C. (1969). *Interpersonal behavior: history and practice of personality theory*. New Jersey: Transaction.

Costa, P. T., & McCrae, R. R. (1987). Neuroticism, somatic complaints, and disease: is the bark worse than the bite? *Journal of personality*, *55*(2), 299-316.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and the Five Factor Inventory (NEO-FFI)*: Professional Manual, Odessa, Florida: Psychological Assessment Resources Inc.

De Raad, B. (2000). *The Big Five Personality Factors: The psycholexical approach to personality*. Hogrefe & Huber Publishers.

De Raad, B. E., & Perugini, M. E. (2002). *Big five assessment*. Hogrefe & Huber Publishers.

deLearyous Interpersoonlijke Communicatie Training. (n.d.). *Training van interpersoonlijke communicatie door natuurlijke taalinteractie met autonome virtuele karakters*. Retrieved July 16, 2014, from http://delearyous.groept.be/nl/home

Detrick, P., & Chibnall, J. T. (2013). Revised NEO Personality Inventory normative data for police officer selection. *Psychological Services, 10*(4), 372-377.

Eysenck, H. J. & Eysenck, M. W. (1985). *Personality and individual differences: a natural science approach.* New York: Plenum.

Eysenck, H. J. (1967). *The biological basis of personality*. Springfield, IL: Thomas.

Eysenck, H. J. (1997). Personality and experimental psychology: The unification of psychology and the possibility of a paradigm. *Journal of Personality and social Psychology*, 73(6), 1224.

Eysenck, H. J., & Eysenck, S. B. G. (1975). *Manual of the Eysenck Personality Questionnaire (junior and adult)*. Hodder and Stoughton.

Gee, J. P. (2003). *What video games have to teach us about learning and literacy*. New York: Palgrave Macmillan.

Gee, J. P. (2007). Games and learning: Issues, perils and potentials. In: Gee, J. P., ed. *Good video games and good learning: Collected essays on video games, learning and literacy*. New York: Palgrave Macmillan, pp. 129–174.

Germeijs, V., & Verschueren, K. (2011). Indecisiveness and Big Five personality factors: Relationship and specificity. *Personality and Individual Differences, 50*(7), 1023-1028.

Giluk, T. L. (2009). Mindfulness, Big Five personality, and affect: A meta-analysis. *Personality and Individual Differences*, 47(8), 805-811.

Graziano, W. G., & Eisenberg, N. (1997). Agreeableness: A dimension of personality. InR. Hogan, J. Johnson, & S. Briggs (Eds.), *Handbook of personality psychology* (pp. 795-824).

Gudjonsson, G. H. (2002). Who makes a good interviewer? Police interviewing and confessions. In M., Bockstaele (Ed.), Politieverhoor en personality-profiling (pp. 93-102). Brussel: Politeia.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.

Hilbig, B. E. (2008). Individual differences in fast-and-frugal decision making: Neuroticism and the recognition heuristic. *Journal of Research in Personality,42*(6), 1641-1645.

Hill, R., Gratch, J., Marsella, S., Rickel, J., Swartout, W., & Traum, D. (2003) Virtual humans in the mission rehearsal exercise system. *Künstliche Intelligenz,* 4(03), 5–10.

Hoekstra, H. A., Ormel, J. & de Fruyt, F. (1996). *Handleiding NEO persoonlijkheids-vragenlijsten NEO-PI-R en NEO-FFI*. Lisse, Swets Test Services.

Hofstee, W. K., de Raad, B., & Goldberg, L. R. (1992). Integration of the Big Five and circumplex approaches to trait structure. *Journal Of Personality And Social Psychology*, 63(1), 146-163.

Hogan, J., & Ones, S. D. (1998). „Conscientiousness and integrity at work", in Hogan, R., Johnson, J., Briggs, S.(ed.), *Handbook of Personality Psychology*. New York: Academic Press.

IBM (n.d.) *IBM SPSS advanced statistics 20* (pp. 46-67). Retrieved from https://blackboard.utwente.nl/webapps/blackboard/content/listContent.jsp?course_id=_15515_1&content_id=_629714_1

Jensen-Campbell, L. A., & Malcolm, K. T. (2007). The importance of conscientiousness in adolescent interpersonal relationships. *Personality and Social Psychology Bulletin*, 33(3), 368-383.

Jones, R. E., & Melcher, B. H. (1982). Personality and the preference for modes of conflict resolution. *Human Relations*, 35(8), 649– 658.

Jung, C. G., (1923). *Psychological types or the psychology of individuation.* New York: Harcourt.

Kabat-Zinn, J. (1990). *Full catastrophe living: Using the wisdom of your body and mind to face stress, pain and illness*. New York: Delacorte.

Krippendorff, K. (2004), Reliability in content analysis: Some common misconceptions and recommendations, *Human Communication Research* 30(3), pp. 411–433.

Kullback, S.; Leibler, R.A. (1951). On Information and Sufficiency. *Annals of Mathematical Statistics* 22 (1): 79–86.

Kullback, S. (1997). *Information theory and statistics*. Courier Dover Publications.

Leary, T. (1957). *Interpersonal Diagnosis of Personality: Functional Theory and Methodology for Personality Evaluation.* New York: Ronald Press.

Lee, K., & Ashton, M. C. (2006). Further assessment of the HEXACO Personality Inventory: two new facet scales and an observer report form.*Psychological assessment*, *18*(2), 182.

Lee, Y. H., Heeter, C., Magerko, B., & Medler, B. (2012). Gaming mindsets: Implicit theories in serious game learning. *Cyberpsychology, Behavior, and Social Networking*, 15(4), 190-194.

Lorr, M., & McNair, D. M. (1963). An interpersonal behavior circle. *The Journal of Abnormal and Social Psychology*, 67(1), 68.

Lorr, M., & McNair, D. M. (1965). Expansion of the interpersonal behavior circle. *Journal of Personality and Social Psychology*, 2(6), 823.

Lorr, M., Bishop, P. F., & McNair, D. M. (1965). Interpersonal types among psychiatric patients. *Journal of Abnormal Psychology,* 70(6), 468.

Lucas, R. & Diener, E. (2000). Personality and subjective well-being across the life span. In Molfese, V. J. and Molfese, D. L. (eds.), *Temperament and personality development across the life span* (pp. 211-34). Mahwah, NJ: Lawrence Erlbaum.

Matthews, G., Deary, I. J., & Whiteman, M.C. (2009). *Personality traits.* Cambridge: University press.

McCrae, R. R. (1996). Social consequences of experiential openness.*Psychological bulletin*, *120*(3), 323.

McCrae, R. R., & Costa Jr, P. T. (1997). Personality trait structure as a human universal. *American psychologist*, *52*(5), 509.

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology*, *52*(1), 81.

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological review*, *102*(2), 246.

O'Connor, B. P. (2002). A quantitative review of the comprehensiveness of the five-factor model in relation to popular personality inventories. *Assessment*, 9(2), 188-203.

Omarzu, J., & Harvey, J. H. (2012). *Interpersonal Perception and Communication. Encyclopedia of Human Behavior* (2nd ed., pp. 465–471). Elsevier Inc.

Pervin, L. A. (1994). A critical analysis of current trait theory. *Psychological Inquiry, 5*(2), 103-113.

Prensky, M. (2001). *Digital game-based learning*. New York: McGraw-Hill.

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies. *Psychological bulletin*, 132(1), 1.

Roe, A. (1957). Early determinants of vocational choice. *Journal of counseling psychology*, 4(3), 212.

Rosenthal, D. (1983). Development of a measure of conflict style: The Rosenthal–Hautaluoma instrument. Unpublished master's thesis, Colorado State University, Fort Collins.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. *Advances in experimental social psychology*, *10,* 173-220.

Sawyer, B. (2007, September). Serious games: Broadening games impact beyond entertainment. In *Computer Graphics Forum* (Vol. 26, No. 3, pp. xviii-xviii). Blackwell Publishing Ltd.

Schaefer, E. S. (1959). A circumplex model for maternal behavior. *The Journal of Abnormal and Social Psychology*, 59(2), 226.

Schaefer, E. S., & Bayley, N. (1963). Maternal behavior, child behavior, and their intercorrelations from infancy through adolescence. *Monographs of the Society for Research in Child Development,* 1-127.

Schmettow, M. (2013, October). Generalized Linear Models. *Research Methods in Human Factors and Engineering*. Lecture conducted from University of Twente, Twente, Overijssel.

Slater, P. E. (1962). Parental behavior and the personality of the child. *The Journal of genetic psychology*, 101(1), 53-68.

Smets, L. (2011). *Police and personality: a quantitative study on investigative interviewing competences and training* (Doctoral dissertation, Ghent University).

Squire, K. (2003). Video games in education. *Int. J. Intell. Games & Simulation*, 2(1), 49-62.

Tellegen, A. (1985). Structures of mood and personality and their relevance to assessing anxiety, with an emphasis on self-report. In: Tuma A. H., Maser J.D., eds. *Anxienty and the Anxiety Disorders.* Hillsdale: Erlbaum; 681-706.

Thorne, A. (1987). The press of personality: A study of conversations between introverts and extraverts. *Journal of Personality and Social Psychology*, *53*(4), 718.

USCICT. (2010, April 27). *Virtual Humans Project SASO-EN* [Video file]. Retrieved from http://www.youtube.com/watch?v=oOp4XP_ziMw.

Vaassen, F., & Daelemans, W. (2010). Emotion classifications in a serious game for training communication skills.Computational Linguistics in the Netherlands. Utrecht, The Netherlands.

Vaassen, F., & Daelemans, W. (2011, June). Automatic emotion classification for interpersonal communication. In *Proceedings of the 2nd Workshop on Computational*

Approaches to Subjectivity and Sentiment Analysis (pp. 104-110). Association for Computational Linguistics.

Van Amelsvoort, A., Rispens, I., & Grolman, H. (2012). *Handleiding verhoor*. Amsterdam: Stapel & De Koning.

van Dijk, B., & Cremers, M. J. (2013). *Actie is reactie: naar effectieve interactie*. Zaltbommer: Thema.

Van Eck, R. (2006). Digital game-based learning: It's not just the digital natives who are restless. *EDUCAUSE review*, 41(2), 16.

Wagenaar, W. A. (1992) Risk taking and accident causation, in *J. F. Yates (ed.) Risk-taking behaviour*, pp. 257–281. Chichester: John Wiley and Sons Ltd.

Wiggins, J. S., & Trapnell, P. D. (1996). A Dyadic-lnteract/onal Perspective on the five-factor Model. *The five-factor model of personality: Theoretical perspectives*, 88.

Wood, V. F., & Bell, P. A. (2008). Predicting interpersonal conflict resolution styles from personality characteristics. *Personality and Individual Differences*,*45*(2), 126-131.

Appendix

1. 2014001_Vragenlijst_Naam en aanvullende vragen
2. 2014002_NEO FFI aanvullende informatie
3. 2014003_NEO FFI antwoordformulier
4. 2014004_Hiernaast zie je een voorbeeld van de Roos van Leary
5. 2014005_Invloed van persoonlijkheid bij het annoteren van verhoren
6. 2014006_Antwoordformulier annotaties