Catching liars by listening carefully

Catching liars by listening carefully

Promises and challenges for verbal credibility assessment

Cover Selected by freepik
Print Datawyse | Universitaire Pers Maastricht
ISBN 978 94 6159 652 9

Catching liars by listening carefully

Promises and challenges for verbal credibility assessment

PROEFSCHRIFT

Ter verkrijging van de graad van doctor aan de Universiteit Maastricht
op gezag van Rector Magnificus, Prof. dr. Rianne M. Letschert
Volgens het besluit van het College van Decanen,
In het openbaar te verdedigen op vrijdag 27 januari 2017 om 10.00 uur.

door

Glynis Bogaard

Promotoren

Prof. dr. Harald L.G.J. Merckelbach
Prof. dr. Aldert Vrij

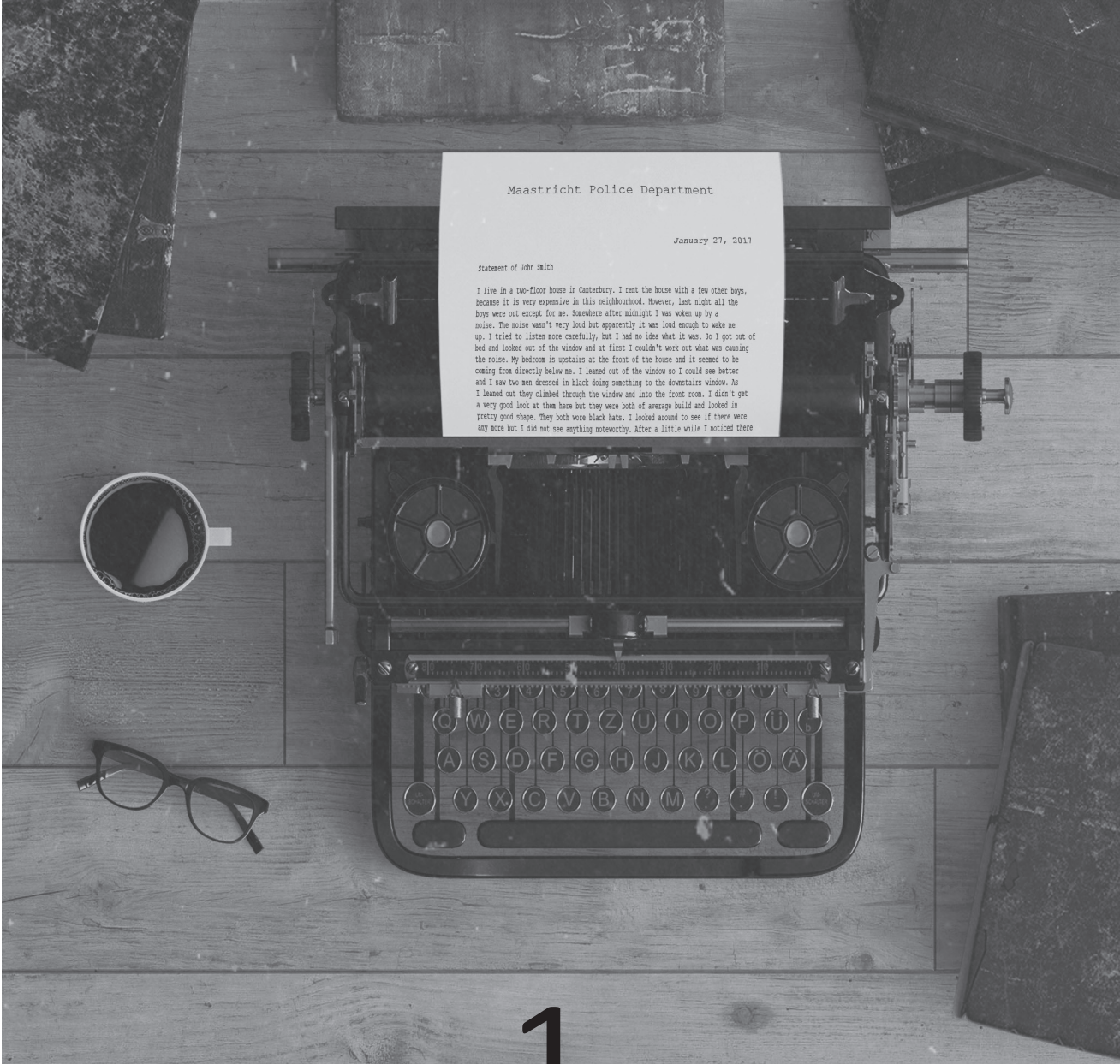
Co-promotor

Dr. Ewout H. Meijer

Beoordelingscommissie:

Prof. dr. Marko Jelicic (voorzitter)
Prof. dr. Ellen Giebels
Prof. dr. Corine de Ruiter
Dr. Miet Vanderhallen
Prof. dr. Geert Vervaeke

CONTENTS



1

General introduction

ABSTRACT

In recent years, Belgian and Dutch police officers have become increasingly interested in verbal credibility assessment methods. These methods gauge the content and linguistic properties of statements to assess their truthfulness. What does psychology have to say about these methods? As will become clear, not only positive things…

This chapter is a translated and extended version of the following articles:

Bogaard, G., Meijer, E., Vrij, A., & Merckelbach, H. (2011). Verbale analysemethoden: Leugenaars praten anders. *De psycholoog*, 10-19.

Bogaard, G., Meijer, E., & Merckelbach, H. (2016). Screenen met SCAN? Liever niet. *Panopticon: Tijdschrift voor Strafrecht, Criminologie en Forensisch Welzijnswerk, 37*, 197-210.

1 INTRODUCTION

In 1995, an American woman named Susan Smith was sentenced to life in prison for the murder of her two children. On October 25 1994, Smith contacted the police saying she had been carjacked by an African-American man, who drove away with her two young sons still in the car. The incident gained a lot of media attention and sympathy from around the world. Yet, a few days later it became clear that Smith fabricated the carjacking story to cover up her actions. She released the brake of her Sedan and let her car roll into South Carolina's Long Lake. Her sons were still strapped into their car seats, meaning she intentionally drowned her children.

When Smith contacted the police, one part of her statement was very remarkable; she used the past tense to talk about her children: "My children wanted me. They needed me. And now I can't help them." Some detectives were alarmed by the use of this past tense. Indeed, one would expect a mother to talk about her missing children in present tense, as long as she believes that her children are still alive. A thorough investigation of the linguistic features of Smith's statement showed she consequently used past tense when talking about her children. By using the past tense in her statements, the police were under the impression that Smith must have known something more about what happened on the 25th of October. Her husband – still convinced that the children were alive – always spoke in present tense when talking about his children (Adams, 1996).

In this example, detectives based their judgment of truthfulness on the linguistic features of a statement. Forensic applications that use this principle – the so-called verbal credibility assessment methods - have become increasingly popular. For example, Belgian and Dutch police officers are enthusiastic about the Scientific Content Analysis (SCAN) (Bockstaele, 2008a, 2008b) and expert witnesses apply the Criteria Based Content Analysis in court cases (van Koppen, 2010).

1.1Statement Validity Analysis and CBCA

The interest in verbal veracity assessment methods dates back to the early 1950's, when psychologist Udo Undeutsch developed the so-called *reality criteria*. While interviewing children, who claimed to have been sexually abused, he noticed that true and fabricated statements differed in their form and content (Undeutsch, 1967). Based on these differences, Undeutsch was the first to produce a list of criteria that could aid in determining the credibility of the statements, though others have published similar lists (Arntzen, 1970; Trankell, 1972). These criteria were later adapted and refined by psychologists Max Steller and Günter Köhnken (1989) into a more formal procedure, called Statement Validity Assessment (SVA). Although SVA was originally developed to investigate the veracity of child witness statements, particularly in sexual abuse cases, research has shown the applicability of SVA is not restricted to these conditions, but can

also be used to judge the veracity of adults' statements (Vrij, 2005; Vrij, Akehurst, Soukara, & Bull, 2002).

SVA is a comprehensive procedure that consists of four stages. The first stage is a case file analysis, comprising the investigation of the background information of the interviewee. The second stage is a semi-structured interview of the witness and is investigative in nature, as it provides additional background information as well as information about the verbal and cognitive abilities of the interviewee. The third, and most important stage, is the Criteria Based Content Analysis (CBCA).

CBCA is based upon the hypothesis that a statement derived from memory differs in certain content and quality features from fabricated statements, which is known as the *Undeutsch hypothesis* (Undeutsch, 1967). Köhnken (1996) divided this hypothesis into cognitive and motivational aspects. The cognitive aspects indicate truthfulness, because only when a person experienced an event s/he is able to meet these criteria. For example, people who experienced an event will be more detailed in their description. In contrast, liars are unable to fall back on an event-specific representation in their memory and hence will not be able to come up with as many different details, as they are deemed too difficult to fabricate. Furthermore, the motivational aspects relate to impression management. Truth tellers are thought not to be as concerned as liars to make a credible impression. For example, truth tellers are more likely to make corrections in their story, or include personally unfavourable information, thereby making it more likely that truth tellers meet these motivational criteria.

CBCA consists of a set of 19 different criteria (see Box 1 and Appendix A), all of which are expected to be more frequently present in truthful than in fabricated statements. More precisely, these criteria are divided into four different categories: General characteristics, specific contents, motivational criteria, and offence-specific elements (see Box 1). The first category consists of three criteria that refer to the statement as a whole. For example, it is expected that truth tellers give a more coherent statement and include more details when describing the event. The second category contains criteria 4 to 13 that refer to specific sentences or words that are concerned with the vividness of the statement. For example, different types of details, such as reproductions of conversations (e.g., "My boyfriend asked me 'what is wrong, you look so pale'") or descriptions about the emotions or thoughts the interviewee experienced during the event (e.g., "I felt so scared when he grabbed my purse"). The third category consists of criteria 14 to 18 that refer to how the interviewee presents his or her statement. For example, truth tellers are expected to make more spontaneous corrections (e.g., "He turned left after the gas station, oh no, I mean he turned right") and are more likely to include self-incriminating details (e.g., "I trusted him and let him in my house, I know now that was very naive"). The fourth category consists of only one criterion, which is concerned with the general knowledge people have about the incident they describe. Truth tellers will include elements in their statement that are known by experts to be typical for this type of incident (e.g., grooming tactics in the case of sexual abuse).

The last stage of the SVA consists of an evaluation of the CBCA results by means of the Validity Checklist (VC). With this checklist, alternative interpretations of the CBCA results (e.g., suggestive questioning) are considered. Thus, CBCA scores are influenced, for example, by factors such as age, cognitive ability, verbal development and fantasy proneness (Merckelbach, 2004; Santtila, Roppola, Runtti, & Niemi, 2000; Schelleman-Offermans & Merckelbach, 2010; Vrij, 2005). The more these alternative factors can be rejected, the more likely it becomes that the CBCA score accurately reflects the veracity of a statement (Vrij, 2008a).

Box 1. Overview of CBCA criteria (see Vrij, 2008)
General characteristics: (1) Logical structure, (2) Unstructured production, (3) Quantity of details. *Specific characteristics*: (4) Contextual embedding, (5) Descriptions of interactions, (6) reproduction of conversation, (7) Unexpected complications during the incident, (8) Unusual details, (9) Superfluous details, (10) Accurately reported details misunderstood, (11) Related external associations, (12) Accounts of subjective mental state and (13) Attribution of perpetrator's mental state. *Motivational criteria*: (14) Spontaneous corrections, (15) Admitting lack of memory, (16) Raising doubts about one's own testimony, (17) Self-deprecation, (18) Pardoning the perpetrator, (19) *Details characteristic of the offence*.

As mentioned before, CBCA is the core feature of SVA, but how accurate is this method? Vrij (2005) investigated the first 37 field and lab studies and concluded that the accuracy of the CBCA varies between 55% and 90%, with an average accuracy around 70%. More recently, three meta-analyses have investigated CBCA (Amado, Arce, & Fariña, 2015; Amado, Arce, Fariña, & Vilariño, 2016; Oberlader, Naefgen, Koppehele-Gossel, et al., 2016). Amado et al. (2015) examined the Undeutsch hypothesis for children's truthful and fabricated accounts and found support that CBCA criteria are, indeed, more present in truthful than in fabricated statements. Amado et al. (2016) conducted a similar meta-analysis for adults' statements and results again corroborated the Undeutsch Hypothesis. Oberlader et al. (2016) conducted the most extensive meta-analysis testing the validity of CBCA and reported a large effect size (g = .97) for CBCA in discriminating between truthful and fabricated statements.

Nonetheless, these meta-analyses report an overall error rate of roughly 30%, meaning that despite CBCA's above chance level accuracy, chances of misclassifying statements are high. Furthermore, CBCA as a whole is more reliable that its separate criteria. Despite these findings, it is not exceptional that an expert witness claims a child's statement must be true because it describes, for example, dialogues (Anson, Golding, & Gully, 1993; Gödert, Gamer, Rill, & Vossel, 2005; Horowitz et al., 1997). Relying on a small subset of CBCA criteria to make credibility assessments strongly decreases CBCA's validity. Furthermore, CBCA experts often claim they do not only test the

statements against the different criteria, but also form an opinion about the accused or the victim based on what can be found in their file. As the SVA procedure recommends, CBCA analysts should start their analysis by investigating all the available information about the case. However, whether this procedure improves the CBCA analysis is questionable. Having all this contextual information could potentially contaminate the CBCA scores, as will become clear in Chapter 5.

Vrij and Mann (2006) tested the cognitive and motivational aspects stipulated by Köhnken (1996). Results again supported the Undeutsch hypothesis, but not the underlying rationales. That is, liars did experience more cognitive load (i.e., test of cognitive criteria). This means that liars must think harder about what to say than truth tellers. Liars have to suppress the truth while at the same time coming up with a convincing story. This causes a higher cognitive demand for liars than truth tellers who can rely on their memory. Furthermore, liars controlled their speech more (i.e., test of motivational criteria) than truth tellers. But, contrary to the hypothesis, results showed a positive correlation for cognitive load and CBCA scores and no correlation for speech control. The authors concluded that this failure to reveal the theoretical rationale of CBCA does not mean it cannot be used as a method to discriminate truthful and fabricated statements, but it does have consequences for making accurate predictions about its usefulness in specific situations.

How about inter-rater reliability of the CBCA? Various studies have investigated this issue and revealed high inter-rater agreement for the total CBCA score, while agreement for the separate criteria varied from low to high agreement (Anson et al., 1993; Gödert et al., 2005; Horowitz et al., 1997; Niveau, Lacasa, Berclaz, & Germond, 2015; Roma, San Martini, Sabatello, Tatarelli, & Ferracuti, 2011). Interestingly, the use of CBCA in practice tends to deviate from the way it is described in scientific papers. Van Nierop, Eshof, and Brandt (2006) investigated CBCA in Dutch practice by analysing the interviews and reports written by CBCA experts. Their findings showed that CBCA experts disagree on which CBCA criteria are most important. Moreover, experts tend to adjust the different criteria to fit their needs. Sometimes these adjustments were so drastic that the criteria could no longer be traced back to their original form. As a result, contrary to what the scientific literature suggests, the use of CBCA in practice does not seem to be standardized.

1.2Reality Monitoring

A second credibility assessment method is called Reality Monitoring (RM). Reality Monitoring originally referred to how people decide on whether memories have an external (real experience) or an internal (imagination) source (Johnson & Raye, 1981). Research assumed that memories that originate from actual experiences differ in quality from memories that originate from imagination. Memories from real experiences are expected to contain more sensory, affective, and contextual information, while memories

from imagination are expected to have more cognitive operations. Cognitive operations include descriptions of inferences like "I think it was cold that day because I was wearing a jacket" and "It appeared to me that she didn't know the layout of the building" (Vrij, 2008b).

To measure the specific characteristics of a particular memory, Johnson, Foley, Suengas, and Raye (1988) developed a 39-item Memory Characteristic Questionnaire (MCQ). Their results showed that the MCQ significantly distinguishes between experienced and imagined events when people judged their own memory. Scientists took this method one step further and investigated the ability of the MCQ to evaluate the quality of someone else's memories. Although research is limited and results only showed a weak indication of MCQ's ability to discriminate (Porter, Yuille, & Lehman, 1999; Schooler, Gerhard, & Loftus, 1986; Vrij, 2008b), researchers saw a window of opportunity for deception detection. They assumed that real and imagined memories were not very different from truthful and deceptive statements. According to these scholars, truths can be considered recalls of actual experienced events, whereas lies can be considered recalls of imagined events. Factor analysis of the MCQ resulted in a set of eight RM criteria that could have potential for deception detection, see Box 2 and Appendix A (Sporer, 1997, 2004).

Box 2. Overview of RM criteria (see Vrij, 2008)
(1) Clarity, (2) Perceptual information, (3) Spatial information, (4) Temporal information, (5) Affect, (6) Reconstructability of the story, (7) Realism, (8) Cognitive operations

Seven of these criteria are expected to be more frequently present in true statements, while the criterion "Cognitive operations" is expected to be more present in made-up statements (Johnson & Raye, 1981; Vrij, Mann, & Leal, 2008). The rationale behind the RM approach is that the quality differences between real and made-up memories are also reflected in speech. For example, a memory of a real experience arises from perception and accordingly is expected to include more perceptual (e.g., "I could smell his aftershave"), temporal (e.g., "It happened around 8 p.m.), spatial (e.g., "He pushed me onto the bathroom floor"), and affective information (e.g., "I felt so ashamed afterwards") than memories that originate from imagination. Furthermore, it is assumed that true memories are more vivid, clear, and sharp compared to made-up memories (Sporer, 1997, 2004). A typical RM analysis is similar to that of CBCA. Subjects are interviewed about certain events and the interviews are taped and subsequently transcribed. Next, RM analysts check for the different criteria within the transcripts. RM criteria are also usually scored on a 3-point scale. However, the criterion "Cognitive operations" is scored in reverse, as this criterion is an indication of deceit.

Vrij (2008) investigated the average accuracy of the RM in 10 studies. Results showed that the percentage of correct classifications with RM is similar to that of CBCA, and varies between 61% and 83% with an average accuracy of 69% (Vrij, 2008). Comparable results were reported by Masip, Sporer, Garido, and Herrero (2005). Recently, Oberlader et al. (2016) also investigated the validity of RM and showed that RM is indeed able to discriminate between truthful and fabricated statement above chance level (effect size g = 1.26), which is comparable to CBCA's validity.

Although CBCA and RM show some degree of overlap – the criterion "contextual embedding" measures the same as the two RM criteria "spatial information" and "temporal information" – there are also clear differences. CBCA was originally developed to judge the statements of young children, while RM has its roots in memory research. The study of Vrij, Akehurst, Soukara, and Bull (2004a) investigated the use of CBCA and RM with four different age groups (5-6, 10-11, 14-15 year olds, and college students), and showed that the RM approach is not effective with younger children's statements. On the other hand, RM is not embedded within a protocol that requires the expert to first consult the file. In this way, the method is less time-consuming and expectancy effects of the "seek and you will find" type might be avoided (see Chapter 5).

1.3Scientific Content Analysis (SCAN)

Unlike the name suggests, the Scientific Content Analysis (SCAN; Sapir, 2005) was not developed by scholars, but by former Israeli lieutenant, Avinoam Sapir. Sapir noticed during his career as a polygraph examiner that suspects who told lies differed in their language from suspects who told the truth.

Box 3. Overview of SCAN criteria (see Vrij, 2008)
(1) Denial of allegations, (2) Social introduction, (3) Spontaneous corrections, (4) Lack of conviction or memory, (5) Structure of the statement, (6) Emotions, (7) Objective and subjective time, (8) Out of sequence and extraneous information, (9) Missing information, (10) First person singular, past tense, (11) Pronouns and (12) Change in language.

Unlike CBCA, the SCAN framework emphasizes that the analyst can investigate the statement without any further knowledge about the file. Typically, a SCAN analysis starts with asking the alleged suspect, witness, or victim to write down 'everything that happened' during a critical period of time. Sapir (2005) refers to this as the 'pure version' of the event, produced without any interference from a police officer. Next, this 'pure version' is matched against criteria such as the extent to which there are gaps in the chronology or the extent to which pronouns (e.g., my, him) are avoided (e.g., when the interviewee mentions "the house" instead of "my house"). The list of SCAN criteria

is extensive and no standardized set exists so far. Chapter 3 aims at addressing this issue by investigating the criteria that drive SCAN. However, Box 3, and Appendix A give an indication of the criteria Vrij (2008) has derived from Sapir's manual.

Contrary to CBCA and RM, SCAN largely consists of criteria that check the particular use of language and structure within the statements and is less focused on the actual content. For example, according to Sapir, deviations from first person singular past tense (e.g., "One saw what he did to her") should raise the suspicion that someone is lying. Furthermore, according to Sapir, 20% of the statement should be devoted to information that led up to the critical event, 50% should be about the critical even, and 30% should be about what happened afterwards. Deviations from this format are thought to indicate deception.

Despite the lack of scientific underpinnings, SCAN has been used by police investigators in Belgium, Canada, Israel, Mexico, the Netherlands, Singapore, South-Africa, the United Kingdom, and the United States (Vrij, 2008a). Furthermore, SCAN is also used by Federal Agencies (including the CIA), Military Law Enforcement (including US Army, US Air Force, US Marine), Private Corporations, and Social Services (retrieved from www.lsiscan.com/id29.htm).

Studies investigating the validity of SCAN are scarce. Apart from the empirical research in this dissertation, only five studies investigated SCAN's validity. Two of these studies are field studies (Driscoll, 1994; Smith, 2001) and the remaining three are lab studies (Nahari, Vrij, & Fisher, 2012; Porter & Yuille, 1996; Vanderhallen, Jaspaert, & Vervaeke, 2015). The oldest field study is that of Driscoll (1994). Depending on the amount of forensic evidence, 30 statements were divided into two categories, *apparently accurate* and *deceptive or doubtful* statements. Next, each statement was analysed for the presence or absence of 10 SCAN criteria. Eight of the 11 (73%) apparently accurate statements and 18 of 19 (95%) deceptive statements were correctly classified by SCAN. However, none of the criteria could accurately discriminate between the two categories. Although these results seem encouraging, there are some methodological issues with this study. An important problem is the absence of ground truth for the different statements, which makes it difficult to interpret these results. Furthermore, Driscoll (1994) used actual police reports and SCAN probably already had an influence on the police investigation, which is highly problematic as it leads to circularity (see Iacono, 1991). Moreover, the study has not been subjected to peer review.

In a second field study, governed by the British Home Office, Smith (2001) investigated 27 statements with SCAN. American police officers categorised these statements on the basis of confession evidence as either truthful (4), deceptive (20) or inconclusive (3). Five different groups of assessors were asked to analyse these statements. Three groups had varying levels of SCAN expertise (occasional, infrequent, and experienced users), and two groups consisted of experienced versus newly recruited police officers who did not make use of SCAN. The groups who used SCAN could accurately classify at least 80% of the true statements and 75% of the fabricated statements. The experi-

enced police officers without knowledge of SCAN could also correctly classify 80% of the statements, which did not significantly differ from the accuracy that could be achieved with the help of SCAN. However, the groups that made use of SCAN did outperform the newly recruited police officers. Remarkably, all groups were able to discriminate between truthful and fabricated statements above chance level. Consequently, SCAN does not seem to have an additional value over experienced detectives in testing a statement on its credibility. Again, one has to be very careful when interpreting these results, as this study also suffers from a lack of ground truth, and was not published in a peer reviewed journal. Moreover, it is not clear whether the three inconclusive statements were also taken into consideration and how they might have influenced the results (Armistead, 2011).

One of the first lab studies that investigated SCAN is the peer-reviewed study of Porter and Yuille (1996). They instructed half of the participants to commit a mock crime while the other half performed an innocuous task. Next, participants were asked to give either (1) a truthful alibi, (2) partially deceptive alibi, (3) false alibi or (4) truthful confession. The statements were analysed with three SCAN criteria (structure of the statement; missing information, and first person singular, past tense). The authors found no differences in the different statements regarding the presence of the different SCAN criteria. In a more recent study, Nahari, Vrij, and Fisher (2012) assigned participants to one of three conditions. The concealment liars were asked to perform a mock crime (stealing an exam) and to conceal this criminal activity but to truly report about their non-criminal activities during the last 30 minutes. Outright liars were asked to conceal their criminal and non-criminal activities and to make up a lie about their activities during the last 30 minutes. Innocents were truth tellers that gave a complete account of their activities during the last 30 minutes. The accounts of all groups were analysed with SCAN, and SCAN scores failed to differentiate between truth tellers and either kind of liar.

The last study that investigated SCAN is that of Vanderhallen et al. (2015). They asked students, experienced detectives and SCAN trained detectives, to assess the credibility of four statements. These statements described a traffic accident that the writer of the statement was involved in. The first two groups were asked to make a judgment without the help of SCAN, while the latter group was instructed to rely on SCAN when making their credibility judgment. Both students and experience police officers were better at recognizing truthful statements than trained detectives who relied on SCAN (82% vs. 57%). With regards to the fabricated statements, there was no significant difference between detectives who relied on SCAN (78%) and those who did not (63%), but both groups were more accurate than students (47%). Although detectives with and without SCAN did not differ with regard to their accuracy for the fabricated statements, reliance on SCAN did result in a significantly lower accuracy for truthful statements. This seems to suggest that the usage of SCAN biased towards to a presumption of guilt.

1.4 THIS DISSERTATION

To sum up, the few methodologically sound studies investigating SCAN indicate that SCAN is unable to accurately discriminate between truthful and fabricated statements. Additionally, the scientific underpinning of SCAN is weak, not to say non-existent, and standardisation is lacking. Nonetheless, SCAN is used worldwide as a lie detection tool. Therefore, we believe that research examining SCAN and its separate criteria is important for informing the field about whether SCAN should be applied in police investigations. Furthermore, both the CBCA and RM approach suffer from a high error rate (30%), which makes single-case assessment (drawing conclusions about the truthfulness of one particular case) challenging. Moreover, many factors outside the veracity of the statements influence credibility assessment. Thus, the current dissertation examines boundary conditions and adjustments of the existing credibility assessment frameworks to improve CBCA's and RM's accuracy.

Outline of this dissertation

This dissertation evaluates three verbal credibility assessment methods, namely Scientific Content Analysis (SCAN), Criteria Based Content Analysis (CBCA), and Reality Monitoring (RM). Its aim is twofold: (1) to evaluate the usefulness of SCAN as a lie detection method, this will be discussed in Part 1, and (2) to investigate boundary conditions and possible improvements for all three methods, which will be discussed in Part 2. In doing so, we are able to draw conclusions about SCANs contribution to deception detection practice, and possible adaptations of CBCA and RM to improve their diagnostic accuracy. The research questions below serve as the starting point for the following chapters.

Part 1: Verbal lie detection and SCAN

Chapter 2: What are the perceived verbal and nonverbal indicators of lie detection?
In this chapter, we investigated whether beliefs about verbal and nonverbal indicators of deception differ between lay people and police officers. Furthermore, we examined whether these beliefs were in agreement with objective indicators known from research. By including SCAN items in the questionnaire, we were also able to investigate to what extent these SCAN criteria appeal to practitioners. If these criteria indeed sound appealing to police officers, this could explain their worldwide application.

Chapter 3: What are the criteria that make up SCAN?
SCAN is increasingly being used by investigative authorities to evaluate the credibility of statements made by witnesses, victims, and suspects. Nonetheless, from the literature it is unclear how many SCAN items need to be included to analyse a statement. Thus, we first investigated which criteria primarily drive SCAN, by examining 82 actual state-

ments in which a SCAN analyses was performed. In doing so, we came up with a list of criteria that were used as a starting point for future research on SCAN.

Chapter 4: Can SCAN tell truth from lies?
Research about SCAN is scarce; as said, only five published studies examined the validity of SCAN (Driscoll, 1994; Nahari et al., 2012; Porter & Yuille, 1996; Smith, 2001; Vanderhallen et al., 2015). Given that SCAN is employed worldwide in police investigations, providing support, or the lack thereof, is not trivial (Meijer et al., 2009). Using a data set of 234 statements, the current chapter aimed at extending previous SCAN findings, and to investigate whether the separate SCAN criteria can actually discriminate between truthful and fabricated statements.

Part 2: SCAN and alternative credibility assessment methods

Chapter 5: Are verbal credibility assessment methods sensitive to contextual information?
What would happen when verbal credibility analysts are supplied with extra-domain information about a case? If verbal credibility assessment methods are sensitive to contextual bias, such extra-domain information would influence the analyst's credibility judgment. In that case, not only the verbal quality of the statement would influence the credibility score, but also potentially unsubstantiated information. This might negatively influence the credibility analyses of the statement. In the current chapter, we investigated to what extent CBCA, RM, and SCAN are sensitive to this contextual information.

Chapter 6: Can an example statement help to increase credibility decisions?
Research has shown that the accuracy of verbal credibility methods is sensitive to certain manipulations. For example, research investigating 'coaching' indicates that providing participants with specific information about the rationale and the different criteria influences CBCA and RM results (Caso, Vrij, Mann, & De Leo, 2006; Vrij et al., 2002; Vrij, Akehurst, Soukara, & Bull, 2004b; Vrij, Kneller, & Mann, 2000). In light of the methods' sensitivities to manipulations, we tested whether giving individuals an example of a detailed statement, fulfilling all relevant criteria of the CBCA, RM and SCAN, might increase the accuracy of the investigated methods.

APPENDIX A

Criteria Based Content Analysis

1. *Logical structure*: The statement should be coherent and does not contain logical inconsistencies or contradictions.
2. *Unstructured production*: Information in the statement is presented in a non-chronological order.
3. *Quantity of details*: The statement should be rich in detail and include specific descriptions of place, time persons, objects and events.
4. *Contextual embedding*: The events should be placed in time and locations, and the actions are connected with other daily activities and/or customs.
5. *Descriptions of interactions*: The statement should contain information that interlinks at least the alleged perpetrator and witness.
6. *Reproduction of conversation*: The statement should report parts of the conversation in the original form or the different speakers are recognizable in the reproduced dialogues.
7. *Unexpected complications during the incident*: Refers to elements that are incorporated in the statement that are somewhat unexpected.
8. *Unusual details*: Refers to details of people, objects, or events that are unique, unexpected or surprising but meaningful in the context.
9. *Superfluous details*: Refers to details that relate to the allegations, but are not essential for the accusation.
10. *Accurately reported details misunderstood*: Refers to when the witness mentions details that are beyond his or her comprehension.
11. *Related external associations*: When events are reported in the statement that are not actually part of the alleged offence but are merely related to the offence.
12. *Accounts of subjective mental state*: Refers to the witness describing the development and changes of his or her feelings at the time of the incident.
13. *Attribution of perpetrator's mental state*: Refers to the witness describing the perpetrator's feelings, thoughts, or motives during the incident.
14. *Spontaneous corrections*: This criterion is fulfilled if corrections are made or information is added to material previously provided in the statement without having been prompted by the interviewer.
15. *Admitting lack of memory*: When a witness admits lack of memory by either saying "I don't know" or "I don't remember" or by giving an answer such as "I forgot all about this except for the part when we were in the car".
16. *Raising doubts about one's own testimony*: When the witness indicates that part of his or her description sounds odd, implausible, unlikely, etc.
17. *Self-deprecation*: When the witness mentions personally unfavorable, self-incriminating details.
18. *Pardoning the perpetrator*: When the witness excuses the alleged perpetrator for his or her behavior or fails to blame the alleged perpetrator.

19. *Details characteristic of the offence*: When a witness describes elements that are known by professionals to be typical for this type of crime, but are counterintuitive for the general public.

Reality Monitoring

1. *Clarity*: Refers to the clarity and vividness of the statement.
2. *Perceptual information*: Refers to the presence of sensory information in a statement (sound, smells, tastes, physical sensations, and visual details)
3. *Spatial information*: Refers to information about locations or the spatial arrangement of people and/or objects.
4. *Temporal information*: Refers to information about when the event happened or explicitly describes a sequence of events.
5. *Affect*: Refers to information that describes how the participant felt during the event.
6. *Reconstructability of the story*: Examines whether it is possible to reconstruct the event on the basis of the information given.
7. *Realism*: Examines whether the story is plausible, realistic, and makes sense.
8. *Cognitive operations*: Refers to descriptions of inferences made by the participant at the time of the event.

Scientific Content Analysis

1. *Denial of allegations*: Whether the examinee directly denies the allegation in the statement by stating "I did not…".
2. *Social introduction*: How the persons described in the statement are introduced.
3. *Spontaneous corrections*: Refers to the presence of corrections in the statement such as crossing out what has been written.
4. *Lack of conviction or memory*: This criterion is present when the writer is vague about certain elements in the statement ("I believe..", I think..", "Kind of..") or when the writer reports that he or she cannot remember something.
5. *Structure of the statement*: Refers to the balance of the statement. In a truthful statement 20% is used to describe activities leading up to the event, the next 50% to describe the actual event, and the final 30% to discuss what happened after the event.
6. *Emotions*: Whether there are emotions described in the statement.
7. *Objective and subjective time*: How different time periods are covered in the statement. Objective time refers to the actual duration of events described, whereas subjective time refers to the number of words to describe these events.
8. *Out of sequence and extraneous information*: Whether the statement recounts the events in chronological order and whether there is extraneous information that does not seem relevant.
9. *Missing information*: Refers to phrases in the statement that indicate that some information has been left out (e.g., finally, later).

10. *First person singular, past tense*: Refers to the format in which a statement is written. Deviations from first person singular, past tense can indicate deception.
11. *Pronouns*: Refers to the use of pronouns in the statement. Omitting pronouns suggests reluctance of the writer to commit him/herself to the described actions.
12. *Change in language*: Refers to the change of terminology or vocabulary in the statement. A change in language indicates that something has altered in the mind of the writer.

Part 1

Verbal lie detection and SCAN

2

Strong, but wrong: Lay people's and police officers' beliefs about verbal and nonverbal cues to deception

Abstract

The present study investigated the beliefs of students and police officers about cues to deception. A total of 95 police officers and 104 undergraduate students filled out a questionnaire addressing beliefs about cues to deception. Twenty-eight verbal cues were included in the questionnaire, all extracted from verbal credibility assessment tools (i.e., CBCA, RM, and SCAN). We investigated to what extent beliefs about nonverbal and verbal cues of deception differed between lay people (students) and police officers, and whether these beliefs were in agreement with objective cues known from research. Both students and police officers believed the usual stereotypical, but non-diagnostic (nonverbal) cues such as gaze aversion and increased movement to be indicative of deception. Yet, participants were less inclined to overestimate the relationship between verbal cues and deception and their beliefs fitted better with what we know from research. The implications of these findings for practice are discussed.

Key words: detecting deceit, nonverbal cues, verbal cues, Criteria Based Content Analysis, Reality Monitoring, Scientific Content Analysis

Published as:
Bogaard, G., Meijer, E., Vrij, A., & Merckelbach, H. (2016). Strong but wrong: beliefs about verbal and nonverbal cues to deception. *PLoS ONE*, *11*, e0156615. doi:10.1371/journal.pone.0156615

2.1 INTRODUCTION

Early research suggests we tell on average two lies each day (DePaulo, Kashy, Kirkendol, Wyer, & Epstein, 1996). More recent studies, however, have shown that there are large individual differences in the prevalence of lie telling, with most lies being told by a minority of people (Halevy, Shalvi, & Verschuere, 2013; Levine, Serota, Carey, & Messer, 2013; Serota, Levine, & Boster, 2010). All these studies suggest that everyone has experience with either being lied to, or with telling lies themselves. Yet, despite this personal experience with deception, research has shown that people, including trained police officers, only perform around chance level in detecting deception (Aamodt & Custer, 2006; C. F. Bond & DePaulo, 2006, 2008; Vrij & Mann, 2001).

One possible explanation for the failure to detect deceit is that people often hold incorrect beliefs about which cues are diagnostic of deception. One notable example here is the belief that liars display more gaze aversion. The Global Deception Research Team (2006) investigated the most widespread beliefs about cues indicative of deception, sampling 2320 participants from 58 countries. Over eleven thousand responses were obtained, resulting in 103 distinct beliefs. Gaze aversion was the belief mentioned by most (64%) participants. Comparable results have been obtained by Strömwall and Granhag (2003), who reported that gaze aversion and an increase in body movement were believed to be strong cues of deception among police officers, judges, and prosecutors. Research shows, however, that gaze aversion is not a sign of deceit (DePaulo et al., 2003; Sporer & Schwandt, 2007).

Incorrect beliefs about cues to deception are not confined to gaze aversion. People tend to rely heavily on nonverbal cues when making deception verdicts (for an overview see Vrij, 2008a), regardless of a large body of research showing that deception cannot be reliably inferred from behavior (DePaulo et al., 2003; Sporer & Schwandt, 2007). Studies from the UK, The Netherlands, and Sweden have compared professionals and lay persons' beliefs about cues to deception, including various professions such as police officers, judges, customs officers, prison guards, and immigration officers as professional lie catchers. These studies revealed that professionals typically hold as many (nonverbal) incorrect beliefs about deception as lay people (Akehurst, Köhnken, Vrij, & Bull, 1996; Strömwall & Granhag, 2003; Strömwall, Granhag, & Hartwig, 2004; Vrij, 2008a; Vrij, Akehurst, & Knight, 2006; Vrij & Semin, 1996). Moreover, when tested against the deception literature, both professionals and lay people overestimated the number of cues that are associated with deception (Hartwig & Bond, 2011; Hartwig & Bond, 2014). More recently, Masip and Herrero (2015) asked police officers and community members how lies can be detected. Again, both groups primarily mentioned beliefs about nonverbal cues.

As people tend to rely primarily on nonverbal cues, the verbal content of the message is largely ignored, despite research showing that diagnostic accuracy can be improved when relying on content (Hauch, Sporer, Michael, & Meissner, 2014; Masip,

Alonso, Garrido, & Antón, 2005; Vrij, 2005, 2008c). Additionally, Mann, Vrij, and Bull (2004) reported that good lie detectors relied more on verbal cues, while poor lie detectors relied more on nonverbal cues. Moreover, meta-analytic research reported a higher lie detection accuracy if the training was based on verbal cues compared with nonverbal training (Hauch et al., 2014). Consequently, content should accordingly be favored over behavior (Levine & McCornack, 2014; Masip & Herrero, 2015; Vrij, 2008c). Surprisingly little research has, however, looked at beliefs about such content cues.

Several veracity assessment methods have been developed that rely specifically on the content of a statement, such as Criteria-Based Content Analysis (CBCA; Steller & Köhnken, 1989) and Reality Monitoring (RM; Johnson & Raye, 1981). CBCA is originally based upon the 'Reality Criteria' that were formulated by Undeutsch (1967), but subsequently transformed by Steller and Köhnken (1989) into the method as it is used to date. The CBCA consists of a list of 19 content criteria that are expected to be more present in true compared with fabricated statements. There is indeed evidence that liars generally tell a less coherent story and are less likely to make spontaneous corrections to their story (e.g., "It was about 2 p.m., oh wait, no it was about 4 p.m."). Also, liars describe fewer reproductions of conversations (e.g., "He told me: 'take off your pants, or someone will get hurt'"). Typically, they will include more contradictions and tell their stories in a more chronological order, for example, because they tend to stick to their rehearsed story. Liars are also less likely to admit forgetting certain details about the event. The following would therefore be more expected for truth tellers than liars: "I know he was wearing a dark blue sweater, but I don't remember the color of his pants" (Amado et al., 2015; DePaulo et al., 2003; Vrij, 2005, 2008b). CBCA was originally developed for evaluating children's testimonies in cases of alleged sexual abuse, but several studies have shown that CBCA can also be used for adults, and is not restricted to sexual abuse cases (Akehurst, Köhnken, & Höfer, 2001; Sporer, 1997; Vrij et al., 2002, 2004a). Both field studies (Lamb et al., 1997; Raskin & Esplin, 1991; Roma et al., 2011) and lab studies (Bogaard, Meijer, & Vrij, 2014; Bogaard, Meijer, Vrij, Broers, & Merckelbach, 2014a; Vrij et al., 2002, 2004b) reported that CBCA is able to accurately discriminate between truthful and fabricated statements. Several meta-analytic reviews, and one meta-analysis have shown that the average accuracy rate of CBCA varies around 70% (Amado et al., 2015; Amado et al., 2016; Oberlader et al., 2016; Vrij, 2005).

RM (Johnson & Raye, 1981) originally stems from memory research and was initially employed for evaluating whether a memory originated from a real experience or an imagined event. The rationale is that a memory from a real experience arises from perception and accordingly will contain more sensory, contextual, and affective information than memories that originate from imagination. It is also assumed that memories of real experienced events are more vivid, clear, and sharp than fabricated memories, which are usually vaguer, less concrete, and are more likely to contain cognitive operations (Sporer, 1997, 2004). Scholars have investigated the usefulness of RM as an aid in assessing credibility. A set of RM criteria has been proposed by Sporer (Sporer,

1997) and entails criteria about specific types of details and items pertaining to the realism and clarity of the statement. Support has been found for a number of RM criteria, namely that liars include less perceptual (e.g., smell, taste, sound), spatial (e.g., location), and temporal (e.g., time, duration) information and that the stories of liars are less plausible than truth tellers' stories (DePaulo et al., 2003; Masip et al., 2005). Meta-analytic reviews have shown that the overall accuracy is similar to that of CBCA and varies around 70 % (Masip et al., 2005; Vrij, 2008d). Although both CBCA and RM have a considerable error margin, it is better than the alternative of relying on intuition (i.e., chance level, see C. F. Bond & DePaulo, 2006, 2008).

As said, little research, however, has investigated people's beliefs about the CBCA and RM cues. One example is Akehurst et al. (1996), where police officers' and laypersons' beliefs about 47 nonverbal cues and 17 content related cues were examined, the latter extracted from RM and CBCA (e.g., spatial and temporal information, emotions, description of conversations). However, the main focus of Akehurst et al. (1996) was on investigating people's beliefs about their own and other peoples' deceptive behavior. Therefore, their study does not allow testing the accuracy of participants' beliefs about verbal and nonverbal cues, and how lay people and police may differ in this respect. Vrij et al. (2006) used the same list of cues and asked police officers, social workers, teachers, and the general public about cues to deception, and how these cues might differ depending on the age of the messenger. No differences between groups regarding their beliefs were reported. Again, the focus of the article was not on the accuracy of the separate cues, but on the group and age differences. Recent research has additionally shown that although both police officers and community members report mostly nonverbal cues when asked how lies could be detected, police officers mentioned more verbal cues (Masip & Herrero, 2015).

The present study aimed to replicate and extend the previous findings of police officers' and lay peoples' beliefs about lie detection cues. In contrast to Akehurst et al. (1996) and Vrij, Akehurst et al. (2006), we also explored participants' views about verbal and nonverbal cues via an open-ended question. In this way, participants were permitted an unlimited number of possible answers, were able to clarify their responses, and could mention cues that were not anticipated on the basis of previous literature. This provides us with detailed information about which cues our participants associate with deception, without influencing them in any way. As stated above, people tend to focus on (invalid) nonverbal cues, but little is known about their insight in verbal cues. Therefore, we examined their views about verbal cues further by asking closed-ended questions related to 28 content cues, instead of 17 content cues used in the previously mentioned studies. These content cues were extracted from CBCA and RM. Moreover, in contrast to previous research, this study focused on the correctness of these beliefs. Insight in these cues is helpful as they shed light on how well practitioners are informed about deception research, and about verbal cues in particular. Preferably, their knowledge about the surveyed cues is better than those of undergraduates.

Besides the beliefs about CBCA and RM criteria, we were also interested in the beliefs about criteria derived from Scientific Content Analysis (SCAN). SCAN is a verbal credibility analyses tool that has been developed by former polygraph examiner Sapir (2005). On the basis of his experience as a polygraph examiner he argued that truth tellers and liars differ in their language. Based on these assumed differences, Sapir developed criteria that could be used to identify deception. According to Sapir, his method is widely used in countries around the world (e.g., Australia, Belgium, Canada, Israel, Mexico, UK, US, the Netherlands, Qatar, Singapore, and South Africa), and is also used by Federal agencies, Military law enforcement, private corporations, and social services (retrieved from www.lsiscan.com/id29.htm). For example, the SCAN course is given on an annual basis to Belgian and Dutch police officers (Bockstaele, 2008a, 2008b). Moreover, Vrij (2008b) describes that SCAN was known by most attendees of an international investigative interviewing seminar, and that many practitioners reported to apply it as a lie detection tool.

Despite its widespread use, no research has supported claims of SCAN's diagnostic accuracy and several studies showed that the SCAN criteria could not differentiate between true and fabricated accounts [Bogaard, Meijer, & Vrij, 2014 (see Chapter 6); Bogaard et al., 2014a (see Chapter 5); Bogaard, Meijer, Vrij, & Merckelbach, 2016; Nahari et al., 2012; Vanderhallen et al., 2015]. In a previous study, we investigated whether –in absence of diagnostic accuracy – susceptibility to confirmation bias could serve as an alternative explanation for SCAN's popularity (see Chapter 2; Bogaard et al., 2014a). In the current study, we extend this line of research by including SCAN criteria to investigate whether SCAN's popularity could be explained by the intuitive plausibility of its items. More precisely, we were interested whether the content criteria used in SCAN would fit with the beliefs people hold about these criteria.

To sum up, the current study explored three issues. First, we investigated which beliefs undergraduates and police officers hold about lie detection in general via an open-ended question, and whether these beliefs were in accordance with the deception literature. Second, we explored the beliefs of both groups regarding the 28 content cues via specific questions, and again checked these beliefs against the deception literature. Third, to test whether beliefs can account for the popularity of SCAN, we investigated to what extent the beliefs about SCAN criteria of both groups were in agreement with the hypothesized direction. Given the exploratory nature of our research, we did not form specific hypotheses.

2.2 METHOD

Participants

A total of 199 participants filled out a questionnaire containing items about cues to deception (see below). The sample consisted of 95 police officers (M_{age} = 44, 64 men) and 104 undergraduate students (M_{age} =19, 18 men). The police officers were recruited by approaching as many police stations as possible, both by phone and by mail informing them about our research. Police officers who expressed an interest in participating were asked to contact the experimenters and were send the link to the questionnaire via email. Participants came from different cities across the Netherlands (e.g., Almelo, Deventer, Groningen, Assen, Maastricht, Sittard, Roermond, Eindhoven, Utrecht, Den Haag). Police officers were either detectives or professional interrogators, so they all had experience with conducting interrogations. They reported a mean experience of 22 years (*SD* = 9.72) ranging from 2.5 to 40 years.

The undergraduate students were recruited through flyers and advertisements at our university campus, or via an online participation system of our university. Participants had an average age of 19 years (*SD* = 1.52) and were mainly first and second year psychology students. These students were included, as they had not yet received any information on lie detection or interviewing techniques in their curriculum. Undergraduates received credit points, whereas the police officers did not receive any compensation for their participation. The study was approved by the ethical committee of the Faculty of Psychology and Neuroscience of Maastricht University. Participants read and signed the appropriate informed consent in accordance with the Declaration of Helsinki and were guaranteed that they could resign from the project at any time and without any consequences.

The questionnaire

Participants first read and signed the online version of the informed consent before starting with the questionnaire. After signing the informed consent, they were asked the open-ended question: "What do you think are good cues for detecting lies?" Participants were given unlimited space to respond. Next, they were asked to indicate their opinion about 28 content cues. Twenty-six of these cues were derived from known verbal credibility assessment tools, namely CBCA (Steller & Köhnken, 1989), RM (Johnson & Raye, 1981), and SCAN (Sapir, 2005).

For CBCA, we included 13 items and excluded the items that have shown to be only rarely present in statements and received little empirical support (i.e., related external associations, raising doubts about one's own testimony, self-deprecation, pardoning the perpetrator, and details characteristic of the offence) (Amado et al., 2015; Vrij, 2005). Furthermore, we excluded the item a*ccurately reported details misunderstood* as it is

used primarily for evaluating child statements. For RM, we included all eight items described by Sporer (1997). The list of SCAN is very elaborate, with some sources reporting as many as 28 items (see Bogaard, Meijer, Vrij, Broers, & Merckelbach, 2014b; Chapter 3). For the current study, we included only those criteria that have been shown to be most frequently used in practice (Bogaard et al., 2014b). This resulted in a list of 12 criteria that are also reported in Vrij (2008b). Some of the criteria included in the SCAN are thought to appear more often in deceptive statements, while others are believed to appear more often in truthful statements. As there was CBCA, RM, and SCAN overlap with regard to six items (i.e., spontaneous corrections, lack of memory, emotions, spatial information, temporal information, extraneous information), we included these items only once. Additionally, we included the item "length" in our survey, as research has shown that truthful stories tend to be longer than fabricated ones (DePaulo et al., 2003). Moreover, we have included the item "self-references", which is a combination of two SCAN criteria (i.e., use of pronouns and first person singular, past tense), and has been shown to be diagnostic in previous studies (see for example DePaulo et al., 2003; Newman, Pennebaker, Berry, & Richards, 2003). This resulted in a list of 28 separate items, which were presented in the order listed in Appendix B.

For every item, we gave a short description illustrated by an example. For example, for spatial information we gave the following description "This cue refers to all descriptions about locations or spatial arrangements of people and/or objects (e.g., He was sitting left to his wife)". Next, as in Strömwall and Granhag (2003) and Granhag, Andersson, and Strömwall (2004) participants were asked to indicate their opinion on forced-choice answer scales with four alternatives, for all items. Respondents could choose between two directed (e.g., 'this cue is used less by liars' or 'this cue is used more by liars') and one neutral (e.g., there is no difference between liars and truth tellers regarding this cue'). Furthermore, a 'don't know' alternative was also always available.

Additionally, participants answered questions about their background, function and experience, and whether they ever had training in deception detection. Furthermore, participants were asked to indicate on a 7-point scale ranging from 1 (very poor) to 7 (very good) how well they thought they would perform in detecting deception and how well they knew the literature about lie detection. The questionnaire was administered online via thesistools.com and all participants received a login name and code to complete the questionnaire. To make sure that the participants completed the entire questionnaire, it was built in such way that participants could not skip any questions. Two students were asked to pilot the questionnaire; they needed approximately 30 – 45 minutes to complete all items.

2.3. RESULTS

None of the participants reported ever having received training in lie detection. In the following we present the answers both groups gave to our questions about their understanding of lie detection literature and their skills in detecting deceit. First, in response to the question 'how well do you know the literature about lie detection?', the police officers reported they were not very knowledgeable about the literature (M = 2.78, SD = 1.47), but still, police officers' self-reported knowledge was more extensive than that of students (M = 1.89, SD = 1.03) [$t(197) = 4.94$, $p < 0.001$, d =0.71]. In response to the question 'how well do you think you would perform in detecting deception?', police officers indicated their self-perceived performance to be moderate (M = 3.92, SD = 1.31), which did not differ from the students (M = 4.15, SD = 0.95) [$t(197) = 1.41$, $p = 0.16$, d = -0.20]. Students who indicated they were more knowledgeable about the literature also indicated they were better at detecting deceit ($r(102) = .33$, $p = 0.001$). In contrast, no significant correlation between literature knowledge and deception detection was found for police officers. However, we did find that experience as an officer (in years) was positively correlated to self-reported lie detection skills ($r(93) = .21$, $p = 0.04$).

We first present the results for the open question using an analysis similar to the one reported by The Global Deception Research Team (2006). Next, we present the results for the closed questions using an analysis similar to that of Strömwall and Granhag (2003) who used a similar response format. In the following sections the analyses will be clarified in more detail.

Open question

Coding of answers. In response to the question: "What do you think are good cues for detecting lies?" widely different responses were obtained. To condensate the data, two raters examined all responses and grouped them into two different categories; nonverbal and verbal cues. Within the nonverbal category, responses were assigned to specific categories such as speech characteristic (e.g., response latency, voice pitch), facial behaviors (e.g., blushing, gaze aversion), and body movements (e.g., illustrators, moving feet). For this purpose, the list of 47 categories employed by Akehurst and colleagues (for the complete list see Akehurst et al., 1996; Vrij et al., 2006) was used. The verbal cues were categorized according to the cues listed in Appendix B.

First, inter-rater reliability of the two raters for presence of cues in responses of participants was calculated. We only coded a cue as present when both raters agreed upon its presence, when raters disagreed upon its present, the cue was scored as absent. As can be seen in Table 2.1 and 2.2, percentages often deviated considerably from 50%, which indicates a skewed data set. This is potentially problematic as Kappa is not an informative measure of agreement with highly skewed marginal distributions. In such cases, the reported Kappa value can in fact be very misleading (See for instance

Table 2.1. Percentage of students (n = 104) and police officers (n = 95) mentioning nonverbal cues.

Items	Students			Police officers		
	Percentage Present	Kappa	Percentage agreement	Percentage Present	Kappa	Percentage agreement
Gaze aversion	51.9	0.88	94.2	23.2	0.92	97.9
Nervous body	35.6	0.86	93.3	10.5	0.85	96.8
Sweating	27.9	0.98	99	11.6	0.95	98.9
Body movements	15.4	0.9	97.1	8.4	0.69	93.7
Facial expressions	10.6	0.82	96.2	5.3	0.9	96.8
Blushing	8.7	1	100	10.5	-	100
Stuttering	7.7	0.88	98.1	7.4	0.93	98.9
Self-manipulations	6.7	0.81	97.1	3.2	0.85	98.9
Faltering speech	5.8	0.79	97.1	1.1	0.49	100
Pitch	5.8	0.85	98.1	2.1	1	100
Hand arm finger movements	5.8	0.76	97.1	3.2	0.85	98.9
Repetitions	4.8	0.9	99	2.1	-	97.9
Postural shifts	4.8	0.76	97.1	10.5	0.81	95.8
Speech characteristics	3.8	0.79	98.1	2.1	1	100
Hectic speech	2.9	1	100	1.1	1	97.9
Pupil dilation	2.9	1	100	0	-	100
Smiling	2.9	1	100	0	-	100
Gesticulations	2.9	0.65	97.1	0	-	100
Behavior (not further specified)	1.9	0.8	99	27.4	0.8	91.6
Evasive responses	1.9	1	100	5.3	0.64	94.7
Response latency	1.9	0.8	99	4.2	0.88	98.9
Eye blinks	1.9	1	100	0	-	100
Shaking	1.9	1	100	3.2	1	100
Leg feet movements	1.9	1	100	3.2	1	100
Grammatical errors	1	0.66	99	1.1	1	100
Soft voice	1	1	100	0	-	100
Swallowing	1	1	100	0	-	100
Head movements	1	1	100	0	-	100
Nervous face	1	0.66	99	0	-	98.9
Pauses	0	-	100	1.1	1	100
Twitches	0	-	100	2.1	0.66	100
Variation in facial	0	-	99	1.1	-	98.9
Change in behavior	0	-0.1	98.1	6.3	0.73	95.8

Note. Although we used the 48 items presented in Akehurst et al. (1996) to categorize the answers of our participants, only 33 different items of this list were actually covered within the answers.

Feinstein & Cicchetti, 1990). To overcome the misleading underestimation of Kappa in our dataset, we also included percentage agreement. As can be seen in Table 2.1 and

2.2, for nearly all cues the prevalence is very low, which explains the discrepancy between percentage of observed agreement (high) and the chance corrected agreement of the Kappa statistic (low). As the low Kappa values were always accommodated by high levels of observed agreement, our values can be considered sufficient to continue our analyses.

To our question "What do you think are good cues for detecting lies?", participants gave a total of 443 different responses. For students, the two raters identified 232 nonverbal cues and 20 verbal cues; for police officers, these were 149 nonverbal cues and 42 verbal cues. Thus, on average, 14 percent of the total responses pertained to verbal cues of deception. Looking at the distinct categories displayed in Table 2.1, the most common nonverbal cues about deception mentioned by students were (1) gaze aversion, (2) nervousness, (3) sweating, (4) body movements, and (5) facial expressions (not further specified). For police officers the most common nonverbal cues were (1) behavior (not further specified), (2) gaze aversion, (3) sweating, (4) nervousness, and (5) blushing. Chi-square analyses were used to identify significant differences between groups. For both students and police officers, *gaze aversion* was the most frequently mentioned cue, but students mentioned it more often than police officers, [χ^2 (1, n = 199) = 17.40, $p < 0.001$]). For the remaining cues, the cue *behavior* was mentioned more often by police officers than by students [χ^2 (1, n = 199) = 26.59, $p < 0.001$]), while students mentioned *sweating* [χ^2 (1, n = 199) = 8.22, $p = 0.004$]), and *nervousness* [χ^2 (1, n = 199) = 17.27, $p < 0.001$]) more often than police officers. For *facial expressions, blushing* and *body movements,* no significant differences emerged between groups.

Table 2.2. Percentage of students (n= 104) and police officers (n = 95) mentioning verbal cues.

	Students			Police officers		
Items	Percentage Present	Kappa	Percentage agreement	Percentage Present	Kappa	Percentage agreement
Contradictions	8.7	0.94	99	31.6	0.86	93.7
Quantity of details	10.6	0.83	96.2	3.2	0.74	97.9
Verbal	0	-	98.1	3.2	1	100
Coherence	0	-	99	2.1	1	100
Plausibility	0	-	99	2.1	1	100
Lack of memory	0	-0.1	98.1	1.1	0.49	97.9
Missing information	0	-	100	1.1	1	100

Note. Although we used the 28 items of Appendix B to categorize the answers of our participants, only 7 different items of this list were covered within the answers of our respondents

For both students and police officers, the most common verbal cues were (1) contradictions and (2) quantity of details (see Table 2.2). Students mentioned *quantity of details* more often than police officers [χ^2 (1, n = 199) = 4.18, $p = 0.04$]), while police officers mentioned *contradictions* more often than students [χ^2 (1, n = 199) = 16.56, $p < 0.001$]). Some police officers (3.2 %) said they used verbal cues to detect deception but

did not specify these cues. Thus, police officers and students listed considerably more (four and 11 times, respectively) nonverbal cues than verbal cues as diagnostic cues.

Closed questions

'Don't know answers

Table 2.3 summarizes endorsement percentages in students and police officers. We first investigated to what extent both groups chose the 'don't know' alternative. To allow for Chi-square tests, we first recoded the data such that both directed answers (i.e., less or more) and the neutral answer (i.e., no difference) were coded as '1' and the 'don't know' answer was coded as '0'. Next, we compared the two groups with each other: 27 out of 28 Chi-squares were significant, meaning that for all but one item (i.e., plausibility [χ^2 (1, n = 199) = 2.51, p = 0.11]), police officers were more conservative, and chose the 'don't know' answer significantly more often than student (i.e., Chi-square values ranged between 6.13 and 19.78).

Table 2.3. Percentage of students (n = 104) and police officers (n = 95) who endorsed the answer options.

	Students				Police officers			
Item	-	0	+	?	-	0	+	?
Denial of allegation	17.31	10.58	64.42	7.69	11.58	27.37	29.47	31.58
Social introduction	34.62	18.27	35.58	11.54	27.37	21.05	15.79	35.79
Coherence	67.31	8.65	24.04	.00	41.05	26.32	18.95	13.68
Clarity	44.23	22.12	31.73	1.92	34.74	20.00	23.16	22.11
Spontaneous corrections	47.12	9.62	38.46	4.81	26.32	15.79	35.79	22.11
Lack of memory	42.31	17.31	37.50	2.88	22.11	14.74	45.26	17.89
Contradictions	12.50	10.58	74.04	2.88	14.74	15.79	51.58	17.89
Perceptual information	62.50	15.38	12.50	9.62	60.00	10.53	6.32	23.16
Main event of statement	42.31	10.58	32.69	14.42	49.47	3.16	20.00	27.37
Emotions	60.58	13.46	18.27	7.69	49.47	17.89	7.37	25.26
Quantity of details	40.38	11.54	43.27	4.80	54.74	12.63	15.79	16.84
Spatial information	46.15	22.12	29.81	1.92	50.53	15.79	11.58	22.11
Objective versus subjective time	50.96	11.54	24.04	13.46	36.84	13.68	8.42	41.05
Unstructured production	31.73	21.15	40.38	6.73	25.26	24.21	31.58	18.95
Description of interaction	51.92	19.23	19.23	9.65	45.26	11.58	12.63	30.53
Temporal information	31.73	29.81	32.69	5.81	41.05	21.05	12.63	25.26
Self-references	38.46	14.42	40.38	6.73	33.68	16.84	20.00	29.47
Extraneous information	34.62	3.85	55.77	5.77	24.21	12.63	42.11	21.05
Missing information	34.62	17.31	42.31	5.77	22.11	15.79	34.74	27.37
Reproduction of conversation	57.69	14.42	19.23	8.65	51.58	10.53	7.37	30.53
Reconstructability	37.50	28.85	25.00	8.65	38.95	24.21	15.79	21.05
First pers sing, past tense	31.73	28.85	16.35	23.08	20.00	17.89	6.32	55.79
Use of pronouns	21.15	37.50	25.96	15.38	16.84	18.95	12.63	51.58

	Students				Police officers			
Item	-	0	+	?	-	0	+	?
Unusual details	63.46	8.65	23.08	4.81	53.68	11.58	12.63	22.11
Plausibility	37.50	36.54	18.27	7.69	46.32	28.42	10.53	14.74
Changes in language	39.42	12.50	38.46	9.62	14.70	20.00	22.10	43.20
Length of the statements	25.00	9.62	57.69	7.69	29.47	25.26	16.84	28.42
Cognitive operations	52.88	14.42	20.19	12.50	47.37	12.63	12.63	27.37
Average	41.30	17.10	33.60	7.90	35.30	17.40	20.00	27.30

Note. "-" means less or fewer when lying, "+" indicates more when lying, and "?" means participants indicated "I don't know".

Directional and neutral answers

We analysed the data as has been done previously in studies using a similar response format (Strömwall & Granhag, 2003). We first recoded directional and neutral answers as -1 (i.e., less when lying), 0 (i.e., no difference), and 1 (i.e., more when lying). We excluded the 'don't know' alternative from the following analysis.

Next, we analysed the data with multiple one-sample sign tests (one for every item in the questionnaire) to investigate whether the average mean value of every item was significantly different from 0. In this way, we were able to investigate whether there was a preference for either one of the directional answers. Means and *p*-values for both groups are presented in Table 2.4. To correct for multiple testing, we have adjusted the alpha level to 0.01. Both groups believed that deceptive statements contained more *denials of allegations,* and more *contradictions* than truthful statements. On the other hand, both groups believed that deceptive statements were less *coherent*, contained less *perceptual information*, fewer *descriptions of emotions*, fewer *descriptions of interactions*, and fewer *reproductions of speech*. Moreover, both groups thought that for liars *objective and subjective time* were less in correspondence than for truth tellers.

Students and police officers also believed that liars use less *first person singular past tense* verbs; tell stories that are less *plausible*, with fewer *unusual details*, and fewer *cognitive operations*. Both groups believed that there was no difference between truth tellers and liars concerning the following cues: *Social introduction, clarity, spontaneous corrections, unstructured production, self-references, extraneous information, missing information, use of pronouns*, and *changes in language*.

We also investigated whether police officers and students significantly differed in their preference for the items in the questionnaires. This was done by means of multiple Mann-Whitney tests (for every item separately). Again, to correct for multiple testing, we adjusted the alpha level to 0.01. Although both groups agreed for most of the cues, they significantly differed in their opinion on six cues (see Table 2.4). Students expressed the belief that deceptive statements were *longer* than truthful statements, while police officers thought there was no difference in this respect. For the remaining five cues (i.e., *main event of the statement, quantity of details, temporal and spatial*

details, and reconstructability of the statement), police officers thought they were less present in deceptive statements, while students believed there were no differences between deceptive and truthful statements with regard to these five cues.

Relationship with extant literature

Table 2.4 summarizes which verbal cues are solid cues of deception according to the extant empirical literature. If evidence about a particular item was mixed, the item was denoted with a -, indicating no clear relationship between the verbal characteristic and deception. Items that were exclusively derived from SCAN (i.e., 8 items, see Appendix B) are discussed in the next section.

As is clear from Table 2.4, not all items in our questionnaire are shown to be effective when detecting lies. This section gives a detailed overview of how we decided on diagnosticity. RM items were scored following the results of DePaulo et al. (2003) and Masip, Sporer, Gardio, and Herrero (2005). A RM item was scored as diagnostic if the item was significantly more present in truthful statements compared to fabricated statements (or vice versa for cognitive operations) in more than 65 % of the included studies. CBCA items were scored based on Amado et al. (2015), DePaulo et al. (2003), and Vrij (2005, 2008b). A CBCA item was scored as diagnostic if the item was significantly more present in truthful statements compared with fabricated statements in more than 65% of the studies included in the meta-analyses of Vrij (2005, 2008b) and/or showed an effect size of at least d = 0.50 in Amado et al. (2015), and/or a significant difference in DePaulo et al. (2003). The remaining items - use of self-references and length of statement – were evaluated based on Newman et al. (2003) and DePaulo et al. (2003) and were scored as diagnostic as both reported significant differences between truthful and fabricated statements regarding these items.

Table 2.4. Beliefs about verbal cues of students (n = 104) and police officers (n = 95).

	Students		Police officers		Between groups	Correct answer[1]
Items	Mean	*p*-value	mean	*p*-value	*Z*-value	
Denial of allegation	.51	<.01	.26	.01	-2.58**	-
Social introduction	**.01**	**1.00**	**-.18**	**.12**	-1.30	-
Coherence	**-.43**	**<.01**	**-.26**	**<.01**	-1.90	<
Clarity	**-.13**	**-.17**	**-.15**	**.18**	-.13	-
Spontaneous corrections	-.09	.40	.12	.30	-1.50	<
Lack of memory	-.05	.66	.28	<.01	-2.43	<
Contradictions	**.63**	**<.01**	**.45**	**<.01**	1.84	>
Perceptual information	**-.55**	**<.01**	**-.70**	**<.01**	1.34	<
Main event of statement	**-.11**	**.31**	-.41	<.01	-2.01*	-
Emotions	-.46	<.01	-.56	<.01	-.51	-
Quantity of details	.03	.83	**-.47**	**<.01**	-3.56**	<
Spatial information	-.17	.07	**-.50**	**<.01**	2.58*	<

	Students		Police officers		Between groups	Correct answer[1]
Items	Mean	*p*-value	mean	*p*-value	*Z*-value	
Objective versus subjective time	-.31	<.01	-.48	<.01	.95	-
Unstructured production	.09	.36	.08	.50	-.15	>
Description of interaction	-.36	<.01	-.47	<.01	-.91	-
Temporal information	.01	1.00	**-.38**	**<.01**	-3.04*	<
Self-references	.02	.91	-.19	.09	-1.47	<
Extraneous information	.22	.03	.23	.04	-.16	-
Missing information	**.08**	**.43**	**.17**	**.13**	-.63	-
Reproduction of conversation	**-.42**	**<.01**	**-.64**	**<.01**	-1.64	<
Reconstructability	**-.14**	**.14**	-.29	<.01	-1.24	-
First person singular, past tense	-.20	.03	-.31	.02	-.70	-
Use of pronouns	.06	.57	-.09	.57	-1.04	<
Unusual details	-.42	<.01	-.53	<.01	-.60	-
Plausibility	**-.21**	**.01**	**-.42**	**<.01**	-1.92	<
Changes in language	**-.01**	**1.00**	**.13**	**.31**	-.90	-
Length of the statements	.35	<.01	-.18	.10	-4.00**	<
Cognitive operations	-.37	<.01	-.48	<.01	-.74	-

Note. Minus sign indicates that the specific criterion is less present for liars. > Verbal characteristics occurs more frequently in deceptive statements; '- No relationship between the verbal characteristic and lying/truth telling; < Verbal characteristics occurs less frequently in deceptive statements. *p<0.01., **p<0.001. Beliefs in the correct direction are in bold. [1]According to DePaulo et al. (2003), Masip et al. (2005), Nahari et al. (2012), Newman and Pennebaker (2003) and Vrij (2005, 2008b).

Evaluation of diagnostic cues - Five out of these 13 diagnostic cues were correctly judged by both students and police officers, namely *coherence, contradictions, perceptual information, reproduction of conversation, and plausibility*. Police officers additionally judged *quantity of details* and *spatial* and *temporal information* correctly (Table 2.5). In sum, as a group, students evaluated five out of 13 diagnostic cues (38%) correctly, while police officers judged eight out of 13 cues (62%) correctly.

For two cues, the groups held beliefs that ran counter to the empirical database; students believed liars gave *longer* statements and police officers thought liars were more likely to admit a *lack of memory*. For the remainder of these 13 diagnostic cues, both groups believed these criteria were not related to deception (i.e., seven for students, four for police officers), yet based on deception literature, there is every reason to assume that they are.

Evaluation of non-diagnostic cues - This leaves us with seven cues that have no basis in empirical evidence, as can be seen in Table 2.4. For the items *clarity*, and *extraneous information* both groups correctly indicated truth tellers and liars did not differ regarding this item. Students also correctly judged *reconstructability* was not diagnostic to detect deceit. For all other items, both groups held the belief that these cues were

diagnostic. We investigated to what extent these opinions were in agreement with the CBCA and RM hypotheses.

For three of them (i.e., emotions, descriptions of interactions, and unusual details), students and police officers indeed followed the hypothetical direction of the CBCA and RM instruments (i.e., criteria are less present when lying). Additionally, police officers' belief about *reconstructability* also followed the hypothesis of RM (i.e., less when lying). For two items both groups mistakenly believed liars included fewer *cognitive operations,* which is contrary to RM and CBCA's hypotheses. Table 2.5 gives a detailed overview of the overall correctness of both groups.

Lastly, we compared the average beliefs about the diagnostic items to those of the non-diagnostic items. We recoded the scores for all the items: the correct answer was coded as 1, the incorrect answer(s) as 0. Next, we calculated the mean scores for the 13 diagnostic, and the seven non-diagnostic items. Two paired samples t-test on these means showed that beliefs about the diagnostic cues were more correct than about the non-diagnostic cues: students judged the diagnostic items (M = 0.40, SD = 0.17) higher than the non-diagnostic items (M = 0.16, SD = 0.18) [t(103)=8.79, p< 0.001, d=1.32], and so did police officers (M = 0.38, SD = 0.21 vs. M = 0.16, SD = 0.22) [t(94)=6.68, p< 0.001, d=1.03].

Table 2.5. Detailed overview of the empirical merits of the beliefs of students (n= 104) and police officers (n = 95).

	Number of items	
Possible outcomes	Students	Police officers
Diagnostic cues (13 items)		
Correct	5 (38%)	8 (62%)
Incorrect	8 (62%)	5 (38%)
Non-diagnostic cues (7 items)		
Correct	2 (29%)	1 (14%)
As hypothesized	3 (42%)	4 (57%)
Incorrect	2 (29%)	2 (29%)
Total correct	7 (35%)	9 (45%)
Total correct and as hypothesized	10 (50%)	13 (65%)

Relationship with SCAN hypotheses

According to the SCAN hypotheses, liars are less likely to directly deny the crime (e.g., they will try to divert from the topic), will fail to correctly introduce persons in their statements (e.g., a correct introduction includes name and role "my son Alex"), and will try to keep the description of the critical event as short as possible. Moreover, the objective and subjective time of their story will not correspond as well as with truth tellers, liars will have more information missing in their stories, use fewer pronouns, and include more changes in language. Nahari et al. (2012) reported no significant differences for any of the SCAN criteria, but Newman and colleagues (2003) found evidence that

liars included fewer pronouns in their statements. As a result, we know that all of the SCAN criteria, except for pronouns, lack diagnostic accuracy. Therefore, we were more interested to what extent the beliefs of both groups were in agreement with the hypothesized direction of SCAN. In this way, we were able to investigate how intuitively appealing the SCAN items are.

For two out of eight criteria - *objective and subjective time* and *first person singular past tense* - both groups agreed with the hypotheses of SCAN. Police officers additionally agreed with the hypothesis that liars describe the *main event* in less detail than truth tellers, while students believed there was no difference between liars and truth tellers in this respect. Note, that there is no empirical evidence to back up these criteria.

For four criteria - *social introduction, missing information, use of pronouns,* and *change in language* - both groups thought that there exists no difference between truth tellers and liars. For one criterion, *denial of allegations*, both groups believed that liars are more likely to deny the allegations, when in fact SCAN's hypothesis states the opposite. The only item of the SCAN list that has been shown to be useful for detecting deception is *use of pronouns*; nonetheless, both groups believed it was not helpful. In total, students followed the hypothesized direction for two out of the eight SCAN items (25%) and police officers for three items (38%).

Correlational evidence for number of correct items

Finally, we investigated whether the number of correctly judged items correlated with our participants' self-reported lie detection skills, knowledge on literature and years of experiences. For police officers, years of experience ($r(93) = .213$, $p = .038$) and self-reported lie detection skills ($r(93) = .266$, $p = .009$) positively correlated with the number of correct answers on the questionnaire. For students, only their self-reported knowledge on the literature positively correlated with the number of correct answers ($r(93) = .214$, $p = .029$).

2.4 DISCUSSION

The current study investigated the beliefs that students and police officers hold about deception. It expands on the extant literature by investigating an extensive list of verbal cues rather than focusing solely on nonverbal cues to deception. Three important issues were explored in this study.

When students and police officers were given the opportunity to list the cues they believed are indicative of deception, they predominantly listed the stereotypical, and unsupported, nonverbal cues (e.g., gaze aversion, nervousness, movement and sweating). These results are in agreement with previous findings (Akehurst et al., 1996; Granhag et al., 2004; Strömwall & Granhag, 2003; Strömwall et al., 2004; Taylor & Hick, 2007; Vrij, 2008a; Vrij et al., 2006; Vrij & Semin, 1996), but they also replicate more

recent results (Masip & Herrero, 2015). In addition, both groups listed considerably more nonverbal than verbal cues as diagnostic cues. This is in line with studies showing that people tend to focus more on nonverbal cues than on verbal cues (Akehurst et al., 1996; Granhag et al., 2004; Strömwall & Granhag, 2003; Vrij et al., 2006), even though the latter are more diagnostic cues to deceit (for a review see Vrij, 2008c).

Two important differences emerged between police officers and students. First, in response to the open question, police officers reported overall less cues than students (191 vs. 252). Second, police officers mentioned more verbal cues than student (42 vs. 20 cues). This finding partially contradicts research by Masip and Herrero (2015), who found that police officers overall reported more cues than lay people. Their data were derived from 22 Spanish officers who were asked to participate during a workshop on eyewitness psychology at their police department, and compared with the answers of 22 community members who were tested in public areas in the same town. Although several differences between policing in Spain and the Netherlands may account for this difference, one notable explanation could be that Dutch police officers are informed during their interrogation training to refrain from making credibility judgments based on behavioral signs (van Amelsvoort, Rispens, & Grolman, 2015). As such they may be weary of nonverbal cues, and list fewer of them.

In terms of nonverbal cues of deception, our findings that especially gaze aversion, nervousness, movements, and sweating were reported as cues to deception, fit with the mistakenly held assumption that liars are more anxious/nervous than truth tellers (Vrij, Granhag, & Porter, 2010). These cues are even reported outside the legal field. Within healthcare professions, nurses as well as therapists have also been shown to hold these false beliefs about deceptive cues (Curtis, 2015; Curtis & Hart, 2015). Moreover, Hart, Hudson, Fillmore, and Griffith (2006) compared managers and non-managers' beliefs about deception cues on the work floor and reported similar results. This incorrect, and widespread assumption might be a result of the common view that lying is bad (The Global Deception Research Team, 2006), and that liars should therefore feel afraid of getting caught. By this understanding, gaze aversion and increases in body movement signal the nervousness that liars feel about their moral dilemma. Most of the behavioral cues that have been mentioned by our participants can be traced back to the idea that lying causes liars to feel distressed, and that this distress is shown in their facial expressions (i.e., blushing, sweating, blinking) or their body (movements, fidgeting, illustrators). However, people seem to underestimate the importance of situational factors that might influence someone's behavior. For example, truth tellers can also be nervous for other reasons than deceptiveness, such as an accusatory interviewing style, the fear of not being believed, or the mere fact of being accused of a criminal act may result in nervous behavior (Ofshe & Leo, 1997).

The second part of our survey investigated beliefs about 28 specific content cues. Students and police officers were largely in agreement about the diagnosticity for most of the listed cues (i.e., 21 cues). For the directional questions, police officers more often

chose the "don't know" answer alternative than students (25% vs. 8%). This suggests that police officers adopt a more conservative threshold for cues to deceit than students. Although many reasons might account for this finding, the most likely explanation is that officers were more concerned with making mistakes than undergraduates. The last decade, much research has focused on investigating police practices, and many of those studies critiqued current practices [e.g., research on interrogation tactics and false confessions, see (Kassin, Appleby, & Perillo, 2010; Kassin et al., 2010)]. Officers' awareness that they were participating in scientific research possibly made them more hesitant to choose a directed response, minimizing their chances of making mistakes.

Interestingly, Strömwall and Granhag (2003) reported a lower percentage of "I don't know" answers for police officers (i.e., 10%) than the current study. One potential explanation for this difference is that we, unlike Strömwall and Granhag (2003), tested participants' beliefs on specific content cues. Participants may have simply been less familiar with these cues, resulting in an increased percentage of 'don't know' answers. Importantly, the relative high percentage of don't know answers for police officers in our study means the results pertaining to the closed questions only reflect the participants who gave a directional answer. On average this amounts to 75% of the police officers, yet for some criteria (i.e., first person singular past tense and use of pronouns) this reflects less than half of the sample.

For the SCAN items, both groups reported beliefs in accordance to the hypothesized direction only for two items, meaning our data do not support our hypotheses that its intuitively appealing items explains SCAN's popularity. Note that previous research failed to support SCAN's diagnostic accuracy, so its popularity cannot be attributed to its accuracy (Bogaard, Meijer, & Vrij, 2014; Bogaard et al., 2014a, 2014b; Nahari et al., 2012). However, our results might be influenced by the way we presented the SCAN criteria. For every criterion included in our questionnaire we gave a description to explain the criterion and we provided participants with an example of one or two sentences. However, in the SCAN manual, the criteria and their interpretation are only vaguely described (Sapir, 2005). Research has shown inter-rater consistency to be low for SCAN in the field, suggesting practitioners adapt the criteria to their own needs (Bogaard et al., 2014b; Smith, 2001). It might be precisely the lack of clear guidelines for the criteria that appeals to practitioners.

We also investigated whether beliefs about the verbal cues were in agreement with the empirical deception literature. Excluding the SCAN items, results revealed that both groups only showed an opposite belief (i.e., less vs. more) for three items. This means that for most of the items participants judged them correctly, or judged them as is hypothesized by CBCA and RM. This was confirmed by the analysis were the diagnostic and non-diagnostic cues were clustered. This finding may explain why untrained lie catchers relying on content cues tend to reach higher accuracy scores than do lie catchers relying on nonverbal cues (Burgoon, Blair, & Strom, 2008; Maier & Thurber, 1968; Mann, Vrij, Fisher, & Robinson, 2008).

Our results add to a body of evidence showing peoples' beliefs in behavioral cues are not indicative of deception. But why are these incorrect beliefs so persistent? Besides reasons that are common in legal psychology, such as illusory correlations and confirmation bias (Hill, Memon, & McGeorge, 2008; Jones & Sugden, 2001; Kassin, Goldstein, & Savitsky, 2003; Nickerson, 1998), two other reasons are particularly important here. First, people - and especially police officers - usually receive inadequate or delayed feedback concerning their credibility judgments (Vrij et al., 2010). That is, for feedback to be effective, an officer should be informed about the truthfulness of the suspect directly after each interview. This does not happen in real life. However, adequate feedback is essential for learning, because people can adjust their decision-making strategy accordingly (Porter, Woodworth, & Birt, 2000). Indeed, lie detection training has been shown most efficient when the training combines information, practice with examples, and feedback (J. E. Driskell, 2012). Without feedback, people are not able to learn their nonverbal beliefs are generally wrong.

Second, police manuals often report subjective ideas on cues to deception instead of relying on scientific research (Vrij et al., 2010). Although this is particularly the case for manuals used in the US, many of the non-diagnostic cues included in these manuals have found their way into popular media (e.g., TV-series and movies). Therefore, people's opinions can be contaminated by the incorrect message that is conveyed about useful deception cues, as is the case for criminal profiling (Snook, Cullen, Bennell, Taylor, & Gendreau, 2008). More precisely, people rely on anecdotal evidence showing that liars display these stereotypical behaviors. Also, repetition of the message that deception can be detected by relying on stereotypical cues strengthens the illusion. Moreover, people often tend to accept information that is conversed to them by presumed experts. As such, if information on these cues can be found in police manuals, people will accept this as evidence for their accuracy.

Two important limitations of our study deserve some attention. First, as we made use of a questionnaire, we can never be sure that participants understood all the items that were included, or what their reasoning was for choosing a specific answer. By providing additional information and an example, we have tried to minimize this issue. Second, we only investigated beliefs and did not look at actual deception detection performance, leaving open the question whether participants with empirically grounded beliefs are better lie detectors (Hartwig & Bond, 2014). Peoples' self-reports about deception tactics do not always correspond with their actual decision making judgments. Nevertheless, there is vast literature showing that people are generally poor at lie detection (Aamodt & Custer, 2006; C. F. Bond & DePaulo, 2006) and one reason for this is that valid cues to deception are rare. Importantly, even the best discriminating nonverbal cues correlate only modestly to deception (DePaulo et al., 2003; Hartwig & Bond, 2011). Furthermore, people tend to strongly rely on nonverbal cues during deception detection (Masip & Herrero, 2015). Our finding that participants had less stereotypical beliefs about verbal cues to deception, might therefore explain the increased

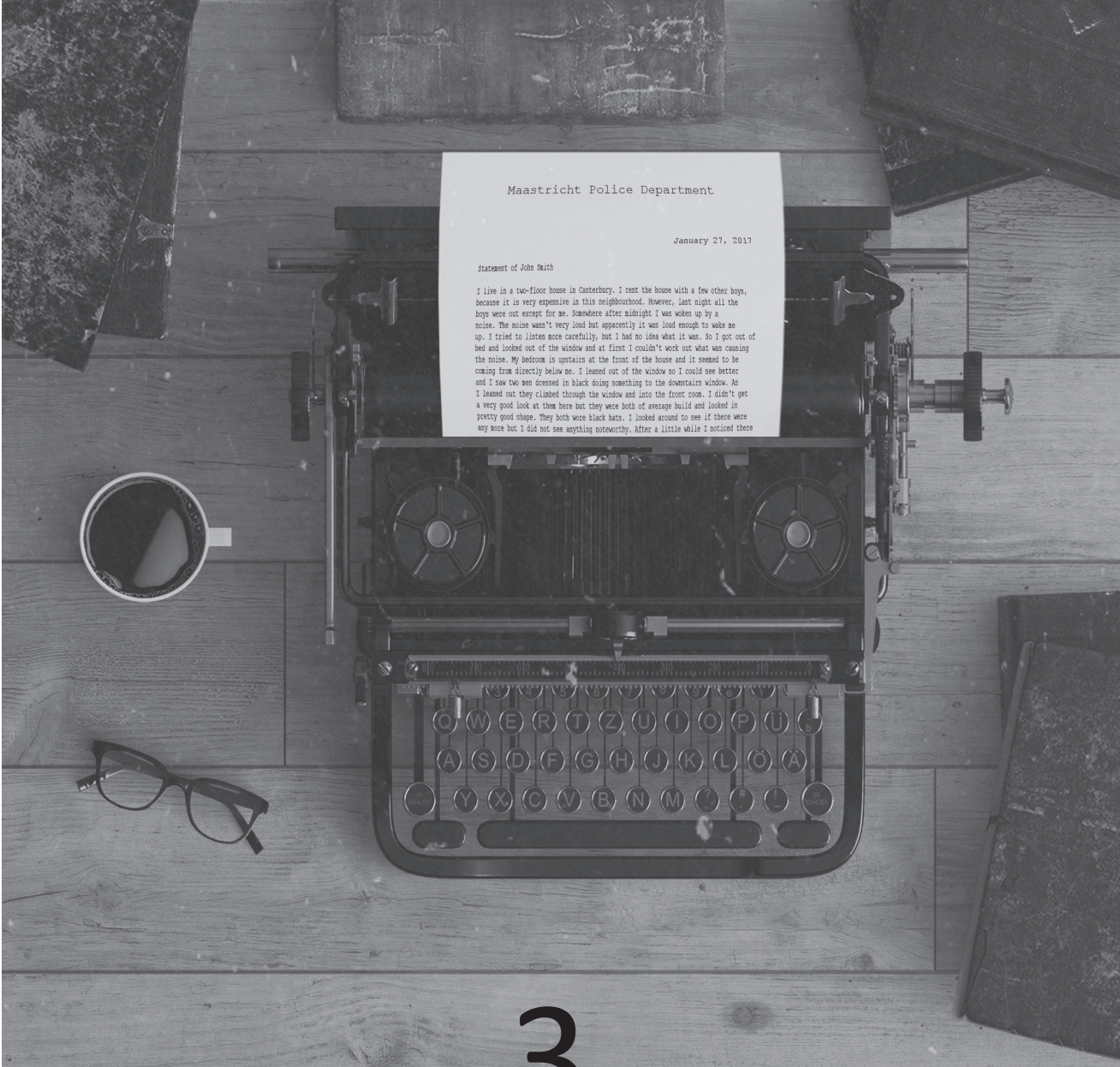
accuracy levels for lie detection when behavioral cues are actively excluded (Davis, Markus, & Walters, 2006; Maier & Thurber, 1968; Mann, Vrij, & Bull, 2002). In any case, whether individuals who hold correct beliefs about verbal cues are better lie detectors is an issue that warrants future research.

In most interrogation settings, police officers have access to visual, vocal, and verbal cues and they may reason that the more cues they can rely upon, the better their detection levels. Furthermore, restricting the presentation mode of suspect statements such that nonverbal cues are excluded, e.g., by focussing on (verbatim) transcripts, takes time, but this is often an issue for police officers. Even so, our results suggest that the mere instruction to attend to verbal cues might increase lie detection accuracy in a naturalistic setting. Research has already shown that training people in how to use content cues increases their detection levels more than training them in nonverbal cues (Hauch et al., 2014). However, for most of the studies included in Hauch et al.'s meta-analysis, statements were presented in the form of transcripts, thereby automatically excluding nonverbal and vocal cues. Consequently, future studies should investigate whether these strong nonverbal cues can be ignored during deception detection, solely by giving the instruction to do so.

In sum, our data demonstrate that both police officers and laypersons hold many incorrect beliefs about the diagnosticity of nonverbal cues, but were less inclined to overestimate the relationship between verbal cues and deception. Here, beliefs fitted better with what we know from research. Although various studies have already shown the dangers of relying on stereotypical nonverbal cues, the current study revealed people still believe these cues to be helpful when unmasking liars. For practitioners, these stereotypical beliefs are potentially harmful (e.g., lie bias), therefore, their diagnosticity – or the lack thereof – should, at the very least, be discussed during police training. Officers should be confronted with these mistaken beliefs and informed about more diagnostic cues. Becoming aware of these wrongful beliefs might be enough to shift their attention to verbal cues, about which - according to our findings - beliefs should be more accurate. This study further investigated whether SCAN's intuitively appealing items might explain the popularity of the method, but results indicated no strong endorsement of SCAN items.

APPENDIX B

1. Denial of allegation – SCAN
2. Social introduction – SCAN
3. Coherence - CBCA
4. Clarity – RM
5. Spontaneous corrections – CBCA and SCAN
6. Lack of memory – CBCA and SCAN
7. Contradictions – CBCA
8. Perceptual information – RM
9. Main event of the statement – SCAN
10. Emotions – CBCA, RM and SCAN
11. Quantity of details – CBCA
12. Spatial information – RM and CBCA
13. Objective versus subjective time – SCAN
14. Unstructured production – CBCA
15. Description of interaction – CBCA
16. Temporal information –RM and CBCA
17. Self- references
18. Extraneous information – SCAN and CBCA
19. Missing information – SCAN
20. Reproduction of conversation – CBCA
21. Reconstructability – RM
22. First person singular, past tense – SCAN
23. Use of pronouns – SCAN
24. Unusual details – CBCA
25. Plausibility – RM
26. Changes in language – SCAN
27. Length of the statement
28. Cognitive operations – RM



3

SCAN is largely driven by 12 criteria: Results from sexual abuse statements

ABSTRACT

Scientific Content Analysis (SCAN) is increasingly used by investigative authorities to evaluate the credibility of statements made by witnesses and suspects. SCAN, however, lacks a well-defined list of criteria, and does not involve a standardized scoring system. In the current study, we investigated which SCAN criteria are represented in actual statements. To this end, we analysed 82 sexual abuse cases of the Dutch police in which SCAN had been applied. Two independent coders scored the presence of various SCAN criteria in the (i) written statements from victims, suspects, and witnesses, and the (ii) recommendations for follow-up investigations that were derived from SCAN. Results showed that SCAN is primarily driven by 12 criteria. Results also indicated a low inter-rater agreement for most SCAN criteria, suggesting SCAN is insufficiently developed as a forensic tool. Still, the 12 criteria can be used as a starting point for future research on their psychometric properties.

Published as:
Bogaard, G., Meijer, E., Vrij, A., Broers, N. J., & Merckelbach, H. (2013). SCAN is largely driven by 12 criteria: Results from field data. *Psychology, Crime and Law*, *20*, 430-499. doi: 10.1080/1068316X.2013.793338

3.1 INTRODUCTION

Investigative authorities are often confronted with deceptive suspects, with witnesses who raise dubious claims, and with bogus victims who fabricate traumatic stories (Greer, 2000; Gudjonsson, Sigurdsson, Asgeirsdottir, & Sigfusdottir, 2007; Lisak, Gardinier, Nicksa, & Cote, 2010). Meanwhile, research has shown the detection of deception to be challenging. Experts, including police officers, generally perform only just above chance level when they base their judgements about deceit on the verbal and nonverbal behavior of suspects or witnesses (Aamodt & Custer, 2006; C. F. Bond & DePaulo, 2006; Vrij, 2008c). Although there is good evidence that verbal indicators of deceit are generally more diagnostic than nonverbal indicators, people intuitively believe that invalid – nonverbal - cues indicate deception (Akehurst et al., 1996; Strömwall & Granhag, 2003; Vrij, 2008a; Vrij et al., 2010; see also Chapter 2).

With this in mind, some researchers (e.g., Vrij, 2008c) have argued that to optimize credibility assessment of suspects, witnesses, and victims, investigative authorities should rely on verbal rather than nonverbal cues to deception. One of the tools that is being used for this purpose is Scientific Content Analysis (Sapir, 2005). SCAN was developed in the mid-eighties by former Israeli polygraph examiner Avinoam Sapir. Over the past few years, SCAN has been used by police investigators in Belgium, Canada, Israel, Mexico, the Netherlands, Singapore, South-Africa, the United Kingdom, and the United States (Vrij, 2008b). Furthermore, SCAN is used by Federal Agencies (including CIA), Military Law Enforcement (including US Army, US Air Force, US Marine), Private Corporations, and Social Services (retrieved from www.lsiscan.com/id29.htm). Typically, a SCAN analysis starts with asking the suspect, witness, or alleged victim to write down 'everything that happened' during a critical period of time. Sapir (2005) refers to this as the 'pure version' of the event, produced without any interference from a police officer. Next, this pure version is matched against criteria such as the extent to which there are gaps in the chronology or the extent to which pronouns are avoided. The outcome of this analysis can be used to make a veracity judgement about the statement and/or to guide subsequent interrogations (see for a more detailed description: Vrij, 2008b).

SCAN has been criticized for lacking a well-defined list of criteria, as well as a standardized scoring system. Police officials using SCAN typically attend a three-day course, but, as the first author discovered while attending this course, no checklist of criteria is offered, nor does the SCAN manual contain such a list. The absence of well-defined SCAN criteria is reflected in the literature, as different authors report different sets of criteria when describing SCAN. For example, in his field study, Driscoll (1994) mentions 10 SCAN criteria, the Belgian police superintendent Bockstaele describes either 14 (Bockstaele, 2008a) or 16 criteria (Bockstaele, 2008b), while in her research governed by the UK's Home Office, Smith (2001) reported 12 criteria. SCAN has also been used in the Netherlands during a pilot by the vice squad of the Amsterdam police (AMS). In this pilot, the police employed 21 criteria derived from the SCAN course manuals. Listing all

linguistic features of SCAN that are mentioned in the literature results in as many as 28 distinct criteria (see Table 3.1 and Appendix C). Moreover, different studies rely on different interpretations of the criteria. For example, Smith's definition of "unnecessary connections/missing information" overlaps with two distinct SCAN criteria - namely "connections" and "missing and/or unnecessary links"- mentioned by Driscoll (1994). It is important to note that despite focusing on a range of criteria, none of the above-mentioned studies found any supporting evidence for SCAN as a lie detection tool.

Given the lack of standardization, it is not surprising that the reliability of SCAN is low. Smith (2001) analysed levels of agreement between eight SCAN analysts, divided into three groups based on their SCAN experience (experienced users, occasional users, and infrequent users) in how they applied SCAN criteria to suspect statements. All analysts were given 27 statements from actual cases and asked to analyse these statements with SCAN. The inter-rater reliability was found to be disappointingly low. Specifically, analysts frequently relied on different criteria when coding the same statement. The highest degree of inter-rater consistency was found for the 'pronouns' criterion, yet this criterion still only achieved a 40% level of agreement. That is, the 'pronouns' criterion was scored to be present in 10 out of 27 statements by one or more groups, while in only 4 out of these 10 statements all groups agreed on the presence of the 'pronouns' criterion. Smith's (2001) study showed that analysts judged many statements as deceptive, but their judgments were based on the use of different criteria.

The present study had two aims. First, to examine which of the various SCAN criteria are present in actual sexual abuse cases statements. We analysed the statements made by suspects and alleged victims in 82 sexual abuse cases that were provided to us by the Dutch police. These statements were part of a pilot conducted by the vice squad of the AMS Police in which SCAN was applied to statements written by suspects or alleged victims in criminal cases. We examined which SCAN criteria were identified by the AMS police SCAN analysts in these written statements. To obtain further insight into which criteria are used by Amsterdam police, we examined which of the SCAN criteria were included by the SCAN analysts in their recommendations for follow-up forensic investigations (for details, see method). Although the current study only relied on sexual abuse statements, the results of this study may still benefit research on SCAN because it would allow for a more standardized list of criteria that can be used in future research. Second, we calculated the inter-rater reliability of the AMS police SCAN analysts to obtain insight into the reliability of the SCAN method in practice, as sufficient inter-rater agreement is an important issue for any assessment method. Thus, the current study addresses reliability rather than validity issues related to the SCAN.

Table 3.1. Literature overview of used SCAN criteria and data police AMS

References	Bockstaele		Driscoll	Smith	Nahari et al.	AMS Police
	2008a	2008b	1994	2001	2011	
1. Social introduction #	x	x	-	x	x	x (coded green)
2. First person singular, past tense <	x	x	x	x	x	x (no code)
3. Unimportant information #	-	-	-	-	-	x (coded yellow)
4. Use of pronouns <	x	x	x	x	x	x (circled red)
5. Structure of the statement <	x	x	x	x	x	x (marked in margin)
6. Missing information #	x	x	x	x	x	x (coded pink)
7. Out of sequence information #	x	x	x	x	x	x (no code)
8. Place of emotions <	x	x	x	x	x	x (underlined)
9. Change in language <	x	x	x	x	x	x (words are linked)
10. Resistance during rape <	-	-	-	-	-	x (marked in margin)
11. First sentence #	-	-	-	-	-	x (no code)
12. Order #	-	-	-	-	-	x (marked in margin)
13. Verb leaving <	-	-	-	-	-	x (coded blue)
14. Communication >	-	-	-	-	-	x (coded orange)
15. Objective vs. subjective time <	x	x	-	x	x	x (no code)
16. Extraneous information <	-	-	-	-	x	x (no code)
17. Together with #	-	-	-	-	-	x (coded purple)
18. Details >	x	x	-	-	-	-
19. Unasked explanation #	-	-	-	-	-	x (coded blue)
20. Lack of conviction or memory <	x	x	x	x	x	-
21. Unexpected complications >	x	x	-	-	-	-
22. Denial of allegation >	-	-	x	x	x	-
23. Spontaneous corrections <	x	x	x	x	x	-
24. Sensory perceptions >	x	x	-	-	-	-
25. Unresponsiveness to questions <	-	x	-	-	-	-
26. Avoiding answers <	-	x	-	-	-	-
27. Activities <	-	-	-	-	-	x (underlined)
28. Exact location #	-	-	-	-	-	x (no code)
29. Negative language #	-	-	-	-	-	x (coded yellow)
Total	14	16	10	12	13	21

Note. Criteria described by the authors are indicated with an "x". Criteria marked with '<' indicate deception, criteria marked with '>' indicate truthfulness. When indicated with '#' see Appendix C for further information. For the coding information of AMS police see The Amsterdam Police Pilot on pp. 5-6.

3.2 METHOD

The Amsterdam Police Pilot

In the vice squad of the Amsterdam police conducted a pilot, using SCAN in 115 sexual abuse cases. Alleged victims, witnesses, and suspects were asked to write down their version of what had happened, upon arrival at the police station, before any interrogation or interview. These statements were then submitted to a SCAN analysis (for details see below), and the results of this analysis were used to guide subsequent interrogation. The goal of this pilot was to explore to what extent SCAN might help in determining the credibility of victims, witnesses, and suspects.

The pilot team consisted of four SCAN analysts (three women), who all had completed the basic SCAN course (see www.lsiscan.com) a few months before the pilot started. At the start of the project, none of the SCAN analysts were experienced in employing SCAN in practice. During the pilot project, they frequently consulted the developer of the instrument (i.e., Avinoam Sapir). The police officers who analysed the statements will be referred to as SCAN analysts throughout the entire manuscript.

All statements in the pilot were analysed by two of the four trained SCAN analysts. In a typical SCAN analysis, the criteria in a statement are first colour coded. Table 3.1 gives a detailed overview of how the SCAN analysts coded each criterion within the written statements. For example, every person who is mentioned in the statement is highlighted in green (Appendix C, criterion 1). For some criteria, their presence is simply indicated in the margin. For example, the structure of the statement is written in the margin at the end of the statement (Appendix C, criterion 5). Finally, a subset of the criteria is not coded in the statement at all, and is only described in the report about the statement (e.g., features like first person singular, past tense; Appendix C, criterion 2). Having analysed the material in this way, each analyst wrote a report about the statement. Based on these reports, a final report was produced that combined all the individual analyses and recommendations. Such a final report would include information about specific questions that could be asked during a subsequent interview based on the SCAN analysis. For example, when the SCAN analysis reveals several unasked explanations within the statement during a specific time frame, a possible recommendation within the report could be: “Ask further explicit questions about the specific time frame”. As another example, the writer may not properly introduce an important person within the statement. According to SCAN this could indicate a difficult relationship between the writer and the introduced person, and a recommended question could be: “Tell us about the relationship with [incomplete introduced person]”. Thus, for each case, the SCAN record, as coded by two SCAN analysts, consisted of:

I. an unprocessed copy of the original written statement of the interviewee;
II. copies of the original statement including the coding scheme of the SCAN analysts (2 versions);

III. individual reports about the statement written by each analyst (2 reports);
IV. a final report that combined the individual analyses and recommendations.

Cases included in the current study

Of the 115 cases that were part of the pilot, 11 could not be used because they were incomplete: nine did not contain a SCAN coded statement, one did not contain a SCAN report with recommendations, and one record consisted of unreadable handwriting. Eighty-two statements were collected by asking alleged victims (n = 43), suspects (n = 18), and witnesses (n = 21). The remaining cases did not contain a written statement, but consisted of other material such as letters from victims (n = 6), suspects (n = 7) and witnesses (n = 7), and e-mails from two witnesses. In accordance with the SCAN guidelines that emphasize the pure narrative version as a starting point for analysis, and to keep the set of statements homogeneous, these cases were also omitted. This resulted in a set of 82 cases that were included in the analysis. All cases were sexual abuse cases.

Analysis

Present/absent coding

SCAN criteria in the written statements and in the final report containing the recommendations were coded separately. To identify the presence of the SCAN criteria in the written statements, two independent coders evaluated whether each of the 21 criteria that were used by AMS police were present within the statements (see Table 3.1 and Appendix C for definitions). One coder followed the SCAN basic course and the other coder read the SCAN course manual and was familiar with the SCAN literature and coding scheme. Thus, these coders did not analyse the statements themselves, but merely investigated the coding scheme that the SCAN analysts had produced for each statement. The evaluation by these coders will be referred to as SCAN evaluation. Presence of a criterion in the coding scheme for a statement was coded as '1' and absence as '0'. Criteria that were coded as present had to be either highlighted within the statement or noted in its margin, resulting in a 0/1 coding per criterion, per statement.

To investigate which of the SCAN criteria contributed to recommendations for follow-up investigations or interviews, the same two independent coders coded whether each of the 21 criteria used by AMS police were present ('1') or absent ('0') in either the individual reports or the final report. Again, this resulted in a 0/1 coding per criterion, per statement.

Frequency

The 0/1 coding described above, however, is suboptimal for situations where a criterion is scored by both analysts, but at different locations within the statement. To obtain more detailed data in the presence of SCAN criteria, we also investigated the exact

frequency of the different criteria within the written statements. This frequency analysis was limited to the written statement, because the recommendations are reports about the criteria, but do not contain the criteria themselves. Consequently, the frequency count will have no additional value over the presence (1) or absence (0) coding of the recommendations.

To identify the exact frequency of the SCAN criteria in the written statements, the same independent coders who coded the presence or absence of the criteria also coded how many times each of the 21 criteria employed by AMS police were present within each statement according to each SCAN analyst. Criteria that were coded as present had to be either colour coded within the statement or noted in its margin. Each coder gave a frequency count of each criterion for each line of the written statements. For example, when a SCAN analyst highlighted two social introductions in line 1 and one introduction in line 6, the total frequency of this criterion would be 3. Furthermore, the coding schemes were compared to make sure that when both coders coded the same number of criteria within one line, these criteria were, in fact, identical. This way, inter-rater reliability can be computed. Merely calculating a total count could produce the illusion of inter-rater agreement since a frequency count of 3, as in the example, can be achieved in multiple ways.

Inter-rater reliability of the two coders for the present/absent coding

First, inter-rater reliability of the two independent coders for the present/absent SCAN evaluations in the written statements was calculated. As can be seen in Table 3.2, percentages of 1's of coder 1 and coder 2 often deviate considerably from 50%. This indicates a very skewed data set, which in turn leads to the underestimation of Kappa and Phi. Therefore, we decided to primarily use proportion agreement. Table 3.2 gives a detailed overview of proportion agreement, Phi and Kappa. Proportion agreement was calculated by dividing the number of SCAN evaluations where both coders agreed on the presence or absence of the criterion by the total number of statements. For example, for the *social introduction* criterion, both coders agreed on its presence in 81 out of 82 SCAN evaluations. This results in an inter-rater reliability of 81/82=0.98. Average proportion agreement for the written statements varied from 0.78 to 1 (M=0.93, SD=0.06).

Inter-rater reliability of the two independent coders for the recommendations was calculated in the same way, and varied between 0.81 and 1 (M=0.90, SD=0.05; see Table 3.2). Kappa and Phi were also calculated and are presented in Table 3.2. For both the written statements and the recommendations, the reliability showed to be sufficient to combine the scores for each criterion between coders. A criterion was coded as present when both coders coded the criterion as present and a criterion was coded as absent when one, or both coders coded the criterion as absent.

Table 3.2. Detailed overview of inter-rater reliability of the two independent coders who coded the presence/absence of the criteria in one or both written statements (top panel) and the recommendations (bottom panel) produced by the SCAN analysts.

Written statements						
Criteria	Kappa	Phi	Percentage Agreement	Percentage present coder 1	Percentage present coder 2	McNemar
4. Pronouns	1	1	1	100	100	nc
19. Unasked explanation	0.73	0.73	93.98	83.13	85.54	ns
13. Verb leaving	0.60	0.61	89.16	18.07	14.46	ns
3. Unimportant info	0.26*	0.26*	93.98	95.18	96.39	ns
14. Communication	0.89	0.89	95.18	65.06	67.47	ns
1. Social introduction	nc	nc	98.80	100	98.80	nc
6. Missing info	1	1	1	89.16	89.16	ns
5. Structure	0.76	0.76	87.95	51.81	51.81	ns
9. Change in language	0.87	0.87	93.98	36.14	32.53	ns
29. Negative language	0.39*	0.40*	78.31	71.08	79.52	ns
17. Together with	0.81	0.81	91.57	68.67	65.06	ns
27. Activities	1	1	1	0	0	nc
12. Order	0.21	0.22	92.77	3.61	6.02	ns
8. Emotion	0.63	0.66	89.16	31.33	19.28	0.006
10. Resistance	0.77	0.77	92.77	20.48	18.07	ns
Recommendations						
Criteria	Kappa	Phi	Percentage Agreement	Percentage present coder 1	Percentage present coder 2	McNemar
4. Pronouns	0.78	0.79	92.77	59.04	53.01	ns
7. Out of sequence info	0.52	0.59	93.98	9.64	3.61	ns
19. Unasked explanation	0.85	0.86	93.98	40.96	33.73	ns
13. Verb leaving	0.63	0.68	92.77	14.46	7.23	ns
3. Unimportant info	0.78	0.80	87.95	61.45	49.40	0.004
14. Communication	0.42	0.52	93.98	8.43	2.41	ns
1. Social introduction	0.86	0.86	91.57	72.29	68.67	ns
6. Missing info	0.83	0.84	91.57	48.19	39.76	ns
2. Commitment	0.69	0.72	86.75	74.70	63.86	ns
5. Structure	0.74	0.77	87.95	69.88	57.83	0.002
15. Obj vs subj time	0.06 (ns)	0.09 (ns)	81.93	16.87	3.614	0.007
9. Change in language	0.88	0.89	91.57	36.14	26.51	ns
16. Extraneous info	nc	nc	86.75	13.25	0	nc
11. First sentence	0.55	0.56	81.93	38.55	19.28	<0.001
28. Exact location	0.50	0.50	91.57	13.25	4.82	ns
29. Negative language	0.92	0.92	96.39	32.53	28.92	ns
17. Together with	0.65	0.69	81.93	57.83	39.76	<0.001
27. Activities	1	1	1	0	0	nc
12. Order	0.44	0.53	86.75	21.69	7.23	<0.001
8. Emotion	0.65	0.69	81.93	57.83	39.76	<0.001
10. Resistance	0.63	0.68	86.75	28.92	15.66	0.001

Note. All *p*'s <0.001, *p**'s < 0.05. ns indicates not significant. nc indicates that these results could not be calculated because at least the score of 1 coder was a constant (all 1 or all 0 scores). Kappa and Phi are often

very low due to the skewness of the data; nevertheless, percentage agreement is very high. Numbers of criteria refer to the description of the particular criterion in Appendix C.

Inter-rater reliability of the two coders for the frequency coding

As an index of inter-rater reliability for the frequency coding, Spearman rho was calculated between the frequency counts of the two independent coders. Separate analyses were carried out for SCAN analyst 1 and SCAN analyst 2. As the frequencies were very skewed we also included a detailed overview of Ms, SDs, median, and skewness for both coders for each analyst (see Table 3.3). For SCAN analyst 1, the Spearman rho of the two independent coders for each separate criterion varied from 0.90 to 1, with an average of 0.99 (*SD* = 0.03). Spearman rho of the two independent coders for SCAN coder 2 varied from 0.70 to 1, with an average of 0.98 (*SD* = 0.08)[2]. More in-depth analyses of the percentages (see Table 3.2, McNemar tests), scored by each of the coders showed some significant differences between both coders. Most significant differences were found for criteria described in the recommendations. This is probably due the unstandardized nature of a recommendation report. Each report follows another structure or outline, depending on the SCAN analyst and type of statement. That is, the criteria are not discussed item-by-item in a specific order, but in a rather unstructured continuous text. Despite these few differences, inter-rater reliability showed to be sufficient. Consequently, for the following analyses we averaged the scores between both raters.

Table 3.3. Detailed overview of the frequency coding of the two independent coders for the written statements, separated for each SCAN analyst.

Analyst 1									
Criteria	Coder 1				Coder 2				Coder 1 & 2
	M	SD	Median	Skewness	M	SD	Median	Skewness	Spearmans rho
4.	53.89	39.12	50	1.022	53.9	39.11	50	1.02	1
19.	2.72	3.08	2	1.65	2.72	3.08	2	1.65	1
13.	0.56	1.07	0	3.09	0.56	1.07	0	3.09	1
3.	3.04	3.2	2	1.32	3.05	3.22	2	1.32	1
14.	3.21	4.15	2	1.5	3.21	4.15	2	1.5	1
1.	15.84	14.47	11	1.65	15.86	14.45	11	1.65	1
6.	3.77	4.26	2	1.51	3.77	4.26	2	1.51	1
5.	0.42	0.52	0	0.6	0.42	0.52	0	0.6	1
9.	0.25	0.49	0	1.84	0.25	0.49	0	1.86	1
29.	2.01	2.56	1	2.3	2.02	2.55	1	2.31	0.99
17.	1.64	2.42	0	1.67	1.53	2.29	0	1.83	0.9
27.	0	0	0	0	0	0	0	0	nc
12.	0.01	0.11	0	9	0.01	0.11	0	9	1
8.	0.53	2.49	0	8.27	0.53	2.49	0	8.27	1
10.	0.14	0.44	0	4.28	0.14	0.44	0	4.25	1

	Analyst 2								
Criteria	Coder 1				Coder 2				Coder 1 & 2
	M	SD	Median	Skewness	M	SD	Median	Skewness	Spearman rho
4.	45.51	44.4	34	1.38	45.51	44.4	34	1.38	1
19.	2.18	2.67	1	1.5	2.21	2.72	1	1.51	1
13.	0.45	0.8	0	2.21	0.45	0.8	0	2.21	1
3.	2.57	3.12	2	1.84	2.57	3.12	2	1.84	1
14.	3.32	4.97	1	2.45	3.32	4.97	1	2.45	1
1.	14.37	17.82	8	1.99	14.39	17.84	8	1.99	1
6.	3.59	4.51	2	1.7	3.66	4.62	2	1.79	1
5.	0.41	0.57	0	1.39	0.41	0.57	0	1.39	1
9.	0.28	0.5	0	1.57	0.28	0.5	0	1.57	1
29.	1.65	2.36	1	2.62	1.73	2.41	1	2.43	0.99
17.	1.71	2.94	0	2.31	1.71	2.93	0	2.33	0.99
27.	0	0	0	0	0	0	0	0	nc
12.	0.01	0.11	0	9.06	0.02	0.16	0	6.28	0.7
8.	0.39	2.04	0	8.16	0.39	2.04	0	8.16	1
10.	0.15	0.52	0	4.39	0.15	0.52	0	4.39	1

Note: Numbers of criteria refer to the description of the particular criterion in Appendix A. nc indicates Spearman rho was not calculated because the variables were constant.

In sum, we will first investigate the presence of the different criteria within the written statements and the recommendations. Second, we will analyse whether some criteria are more prominent in either the (i) written statements or the (ii) recommendations. Third, we will investigate the inter-rater reliability between the SCAN analysts to see whether SCAN can be used in a reliable way.

3.3 RESULTS

Based on the present/absent coding we calculated the proportions of written statements and recommendations in which each SCAN criterion was present, which is presented in Figure 1. For the written statements, results indicated that ten of the criteria were present in more than 25% of the statements (denoted with a * in Figure 1). Furthermore, results indicated that 11 criteria were present in more than 25% of the recommendations (denoted with a * in Figure 1). Nine of these criteria overlapped between the written statements and the recommendations. These criteria were: “Use of pronouns”, “Social introduction”, “Unimportant information”, “Missing information”, “Unasked explanation”, “Negative language”, “Together with”, “Structure of the statement”, and “Change in language”. Three of the criteria differed notably between the written statements and the recommendations. The criterion “Communication” was present in over 25% of the written statements, but not in the recommendations. The

criteria “First person singular, past tense” and “Place of emotions” were present in over 25% of the recommendations, but not in the written statements. Considering all criteria that were present in more than 25% of either the written statements or the recommendations results in a list of 12 SCAN criteria.

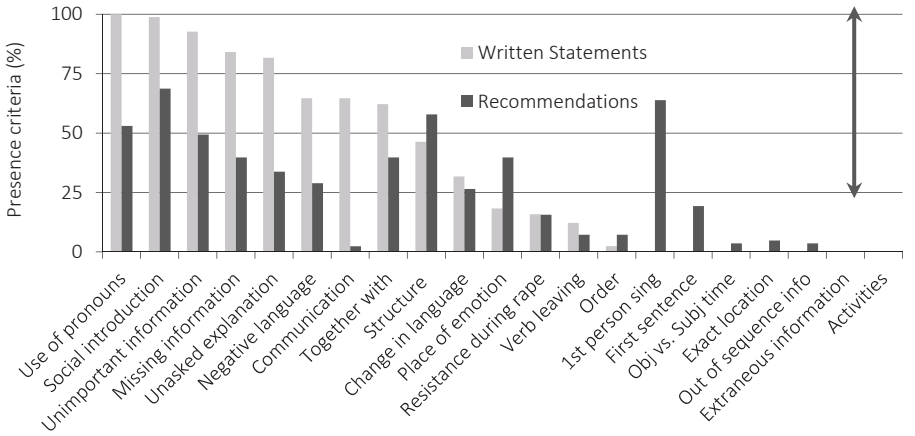


Figure 1. Overview of the SCAN criteria that were present in the written statements, and the criteria that were present in the recommendations for further investigation. Criteria that fall within the range of the arrow (right) were present in more than 25% of the written statements or the recommendations. The X-axis gives the name of each criterion, definitions can be found in Appendix C.

To investigate whether some criteria were more prominent in the written statements or the recommendations, a series of McNemar tests were carried out. To correct for multiple testing, a significance level of 0.01 was used. Only 15 out of the 21 criteria could be scored in the Written Statements, therefore we can only calculate 15 McNemar tests, the remaining six criteria were only discussed in the Recommendations. Nine out of 15 criteria were found to be significantly more often present in the written statements than in the recommendations (see Figure 1). These criteria were “Use of pronouns” [χ^2 (1, N=82) = 36.03, p<0.01], “Unimportant information” [χ^2 (1, N=82) = 28.20, p<0.01], “Together with” [χ^2 (1, N=82) = 9.03, p<0.01], “Missing information” [χ^2 (1, N=82) = 32.24, p<0.01], “Unasked explanation” [χ^2 (1, N=82) = 33.58, p<0.01], “Negative language” [χ^2 (1, N=82) = 18.23, p<0.01], “Communication” [χ^2 (1, N=82) = 49.02, p<0.01], “Place of emotion” [χ^2 (1, N=82) = n.c.[1], p<0.01] and “Social introduction” [χ^2(1, N=82) = n.c., p<0.01. None of the remaining criteria were more often present in the Written Statements compared to the Recommendations. As a result, no differences were found for “Verb leaving”, “Structure of the statement”, “Change in language”, “Order”, “Activities”, and “Resistance”.

[1] N.c. = not calculated. No McNemar statistic could be computed, because one of the variables showed no variation.

Table 3.4. Detailed overview of inter-rater reliability of the SCAN analysts for the written statements

Criteria	Proportion agreement
4. Pronouns	0.79
19. Unasked explanation	0.29
13. Verb leaving	0.32
3. Unimportant information	0.18
14. Communication	0.27
1. Social introduction	0.65
6. Missing info	0.30
5. Structure	0.32
9. Change in language	0.14
29. Negative language	0.30
17. Together with	0.46
27. Activities	0
12. Order	0
8. Emotion	0.49
10. Resistance	0.2
Average	0.31

Note. Only 15 criteria were coded within the statements so inter-rater reliability could only be calculated for these 15. Numbers of criteria refer to the description of the particular criterion in Appendix C.

Next, the inter-rater reliability of the two SCAN analysts was calculated by means of proportion agreement. As the presence of each criterion was coded per line, we could compare the frequency counts of both analysts with great precision. An example will help to clarify how reliability was calculated. Consider SCAN analyst 1 who highlights 3 pronouns in line 1, 2 pronouns in line 2, and 5 pronouns in line 3. SCAN analyst 2, in contrast, highlights 4 pronouns in line 1, 2 pronouns in line 2, and 4 pronouns in line 3. First, we look at the comparison between both analysts for line 1. Here we see that they agree about 3 pronouns. For line 2, they agree about 2 pronouns and for line 3 they agree about 4 pronouns. Statements were also compared to make sure that when both analysts coded the same number of criteria within one line; the criteria were, in fact, identical. This leads to a total amount of 9 pronouns on which both analysts agree. However, in total there are 11 (4+2+5) pronouns scored in the three lines. This leads to a proportion agreement of 9/11= .81. This calculation was carried out for each criterion in each statement. Next, we calculated the average proportion agreement per criterion for all 82 statements (See Table 3.4). Average proportion agreement is 0.31 (*SD* = 0.21). The frequency counts of each criterion were averaged for both analysts to obtain the total frequency count of the different criteria within the written statements, which is presented in Table 3.5.

Table 3.5. Total frequency count for each criterion and frequency separated for the role of the writer, corrected for the length of the statement.

Criteria	Total	M (SD)	Median	Interquartile range	victims 50 lines	suspects 50 lines	witness 50 lines
4. Use of pronouns	4463.50	54.43 (38.79)	52	49.75	55.15	47.29	43.53
1. Social Introduction	1320	16.10 (14.25)	10.5	16.75	12.13	20.39	17.87
6. Missing Information	319	3.89 (3.75)	3	5.38	4.07	4.57	2.11
14. Communication	286	3.49 (4.43)	2	4.50	3.19	3.27	3.38
3. Unimportant information	247.5	3.02 (2.90)	2	3.50	3.01	2.32	2.70
19. Unasked explanation	232	2.83 (2.75)	2	3.88	2.75	3.62	1.81
29. Negative language	167.50	2.04 (2.29)	1.5	2	2.15	0.74	0.89
17. Together with	146	1.78 (2.26)	1	2.88	1.50	2.53	1.48
13. Verb leaving	41.50	0.51 (0.81)	0	1	0.52	0.46	0.38
8. Place of emotion	40	0.49 (2.24)	0	0.50	0.70	0.25	0.06
5. Structure	36.50	0.45 (0.43)	0.5	0.5	0.44	0.50	0.31
9. Change in language	23.50	0.29 (0.41)	0	0.5	0.27	0.39	0.19
10. Resistance	12	0.15 (0.43)	0	0	0.23	0	0.02
12. Order	1.50	0.02 (0.10)	0	0	0	0.11	0
27. Activities	0	0 (0)	0	0	0	0	0

Note. Numbers of criteria refer to the description of the particular criterion in Appendix C.

Finally, we determined these frequencies for victims (n = 43), suspects (n = 18), and witnesses (n = 21), separately. The total frequencies for these categories were analysed by one-way Analyses of Covariance to evaluate differences between the writers in the presence of criteria, with length of the statement as a covariate. To correct for multiple testing, a more conservative significance level of 0.01 was used. However, no significant differences for any of the criteria were found between victims, suspects, and witnesses. Table 3.5 presents the frequencies of the different criteria for each role per 50 lines, the average length of a statement.

3.4 DISCUSSION

Given the unstandardized nature of the SCAN method, the present study examined which SCAN criteria are most frequently used in the field. To address this issue, 82 cases on which AMS police performed SCAN were investigated. Ten criteria were found to be present in over 25% of the written statements and 11 criteria were found to be present in over 25% of the recommendations. Taken together, this resulted in a list of 12 unique criteria that were used in more than 25% of the cases.

These 12 criteria largely overlap with the criteria reported by Smith (2001), and Vrij (2008b) in his review of SCAN. The only criterion that is not listed by these authors is the “together with” criterion, which contributed to the recommendations for follow-up

investigations in 40% of the cases. Two other criteria ("unasked explanation" and "unimportant information") are not directly discussed in Vrij (2008b), but these criteria show some degree of overlap with Vrij's descriptions of the criteria "out of sequence information" and "extraneous information".

Our results also show that 8 criteria were more often present in written statements than in the recommendations, regardless of the examinee's status (i.e., victim, suspect or witness). Two criteria were more often present in the recommendations than in the written statements. These criteria were "First person singular, past tense" and "First sentence". This can be explained by the absence of a coding symbol with which their presence can be marked in written statements. Even though "First person singular, past tense" was not scored in the written statements, it emerged in follow-up decisions in almost 60% of the cases (Figure 1).

To evaluate the criteria used in the written statements and the reliability of the SCAN method, we calculated the exact frequency of each criterion. Although the SCAN analysts were trained in using SCAN, their inter-rater reliability was found to be disappointingly low. This finding is especially striking, as all 4 analysts employed the same definitions – given in a self-produced summary - and should therefore have been able to investigate these criteria and their appropriate interpretations in a similar way. This low inter-rater reliability suggests that the SCAN method is insufficiently developed as a forensic tool.

Several of the most commonly cited SCAN criteria overlap with criteria from other verbal veracity assessment methods. The criterion "place of emotion", for example, overlaps with the criterion "accounts of subjective mental state" of the Criteria Based Content Analysis (Steller & Köhnken, 1989; Undeutsch, 1967), and the criterion "affect" of the Reality Monitoring approach (Johnson & Raye, 1981; Sporer, 1997). However, in SCAN, it is primarily the place of the emotion in the statement that is of importance. More specifically, according to SCAN, truth-tellers include emotions particularly after the climax of the story, while liars tend to mention emotions just before the climax of the story. In contrast, in CBCA and RM the presence of emotional language per se is an indication of authenticity regardless of its precise placement. For example, "I was very scared when he touched me" would fulfil the criterion for CBCA and RM regardless of where this sentence is mentioned within the statement. The SCAN criterion "spontaneous corrections" is similar to CBCA criterion 14, also called spontaneous corrections (for a detailed description of CBCA see Vrij, 2008b), and refers to the presence of corrections within the statement. For SCAN, their presence indicates deceit, while CBCA analysts believe that their presence indicates truthfulness.

A similar discrepancy occurs for the SCAN criterion "lack of conviction or memory", which comes close to CBCA criterion 15 with the same name. CBCA analysts believe that the presence of this criterion indicates truthfulness, while SCAN advocates the opposite belief. The criteria "out of sequence information" and "extraneous information" are similar to the CBCA criteria "unstructured production" and "superfluous details", but

again, both approaches attach opposite qualities to the presence of these criteria. Again, CBCA assumes these criteria make a statement more credible, yet SCAN assumes they designate deception. Research has shown that CBCA criteria are significantly more present in truthful statements relative to false statements (Vrij, 2005). Things are quite different for the related SCAN criteria: recent research has found no evidence that these SCAN criteria are more present in false statements than in truthful statements (Nahari et al., 2012).

Support for certain SCAN criteria comes from research on linguistics features of false and true statements. Newman, Pennebaker, Berry, and Richard (2003) reported that compared with truth-tellers, liars use fewer self-references (e.g., I, me and my) in their stories. These findings were supported by DePaulo and colleagues (2003). By using fewer self-references, liars tend to distance themselves from their statements. The SCAN criteria "first person singular, past tense "and "pronouns" are related to this finding. Sapir (2005) contends that the presence of these two criteria indicates deception, which is in accordance with the findings of the DePaulo et al. (2003) and Newman et al. (2003).

Two limitations of the present dataset deserve attention. First, all the data generated by the vice squad of AMS police were sexual abuse cases. It is possible that some, or even all of the most frequently used criteria are specific for these cases. However, the available literature about SCAN describes similar criteria (Vrij, 2008b; Smith, 2001), which indicates that these criteria are probably also common for non-vice cases. Second, the pilot only included four SCAN trained analysts. Given the unstandardized nature of SCAN, it is possible that these results do not generalize to other SCAN users. Still, the list of most frequently used criteria largely overlaps with those reported in the literature (Vrij, 2008b; Smith, 2001), supporting the notion that these criteria indeed drive SCAN.

In sum, the results of the present study show that SCAN is largely driven by a set of 12 criteria. However, it is important to note that the inter-rater reliability of those criteria is low, except for "Pronouns" and "Social introduction". As sufficient inter-rater reliability is one of the requirements for any tool to be applied in practice, these findings suggest that SCAN is insufficiently developed as a forensic tool and the extensive use of SCAN in practice should be discouraged. Still, to our knowledge, our study is the first to show which criteria drive SCAN. We therefore recommend using the 12 SCAN criteria as a starting point for future studies on their psychometric properties. To overcome the issue of low inter-rater reliability future studies could use a more standardized coding system for scoring SCAN, as was shown by Nahari et al. (2011).

APPENDIX C

Description of SCAN criteria

The SCAN manual does not consist of a clear list of criteria and definition. Instead all criteria, explanations and examples are spread throughout the entire manual. Therefore, we refer to the definitions given by AMS police who studied the manual in detail and extracted the information from the manual.

1. *Social introduction:* This criterion refers to whether persons that are introduced in the statement are introduced by their name and role (e.g. My son, David). When a person in incompletely introduced this could point to a bad relationship between the writer and the introduced person, especially when other persons are introduced correctly (AMS Police; Sapir, 2005, point 236-258, p. 69-73)
2. *First person singular, past tense:* This criterion is also called the test of commitment which states that a truthful person will write his/her statement in the first person singular, past tense. Especially deviations from past tense or writing in third person could indicate a lack of commitment, which, in turn, could indicate deception (AMS Police; Sapir, 2005, point 179-194, p.47-50).
3. *Unimportant information:* This criterion refers to information that has no function in the statement. This means that the statement could be logically understood without this information. The writer did not have to include this information in the statement but did it anyway. Therefore, according to SCAN, this information is very sensitive and important (AMS Police; Sapir, 2005, point 162d, p. 35)
4. *Use of pronouns:* This criterion refers to the use of pronouns in the statement (e.g., "he", "my", "your" etc.). When pronouns are missing within a statement, or more pronouns are expected, this could suggest that the writer wants to distance him/herself from the statement. This could indicate deception (AMS Police; Sapir, 2005, point 260-269, p. 74-77).
5. *Structure of the statement*: In a truthful statement, it is expected that 20% is used to write the prologue (e.g., what happened before the main event), 50% is used to describe the main event, and 30% is used to write the epilogue (e.g., discussion about what happened after the event). Large deviations from this structure could indicate deception (AMS Police; Sapir, 2005, point 444-454, p. 119-123).
6. *Missing information*: This criterion refers to information that is missing and are usually designated by words such as "after a while", "shortly thereafter". "the next thing I remember", etc. It is normal that a person does not tell everything, but missing information situated around the main event could suggest the writer is deliberately hiding important information (AMS Police; Sapir, 2005, point 394-404, p. 110-112).
7. *Out of sequence information*: This criterion refers to information that is given by the writer and has no apparent meaning for the reader. This information feels out

of place in the statement but still is regarded as important information since the writer is giving this information for a reason (AMS Police; point 458-459, p. 124-125).

8. *Place of emotions*: For SCAN it is especially the place of the emotions that is of great importance. Usually emotions will be found in the epilogue of the statement. When emotions are already described during the main event, it is possible that the writer is deceptive (AMS Police)
9. *Change in language*: This criterion refers to a change of terminology or vocabulary in the statement. Especially important are words related to categories such as family members, people, communication, transport or weapons. When a change of language is obvious in a statement (e.g. weapon to gun) and no justification can be found for such a change, this could indicate deception (AMS Police; Sapir, 2005, point 162f-162j, p. 37-38 and point 271, p. 80, and point 461-473, p. 125-127).
10. *Resistance during rape:* With SCAN it is expected that victims of sexual abuse write something about how they tried to resist the offence. When there is no resistance mentioned in the statement this may indicate deception (AMS Police).
11. *First sentence:* According to SCAN the first sentence is a very important sentence in the statement. A lot of information can be found in de first sentence (AMS Police; Sapir, 2005, point 123, p. 27).
12. *Order:* This criterion refers to the order in which persons or objects are mentioned in the statement. In this way the writer reveals his/her priority regarding these persons or objects (AMS Police; Sapir, 2005, point 259, p. 73).
13. *Verb leaving:* According to SCAN the verb leaving is important. Using this term in the statement may indicate deception, especially when this verb is used in the first sentence (AMS Police; Sapir, 2005, point 273-275, p. 80-81).
14. *Communication:* According to SCAN every verb in relation to communication is important. When a writer is able to cite parts of conversations in the statement this indicates truthfulness (AMS Police; Sapir, 2005, point 312, p. 91).
15. *Objective versus subjective time:* This criterion refers to the relationship between subjective and objective time. Subjective time is the pace of the statement; more precisely how many lines are necessary to write about one objective hour. On average it is expected that a writer needs 3 or 4 lines per hour when describing one day. Large deviations from this pace suggest deception (AMS Police, Sapir, 2005, point 280-295, p. 81-86).
16. *Extraneous information:* This criterion refers to information that does not seem relevant or seems strange for the reader. According to SCAN the use of this information may be used to divert attention from other information which can indicate deception (AMS Police, Sapir, 2005, point 173a-173b, p. 42-43)
17. *Together with:* According to SCAN the use of the pronoun “we” indicate that the writer feels a certain commitment to the other person. However, when a writer uses the term “together with” there is a lower sense of commitment to the other

person. This information is used to highlight tension between the different persons mentioned in the statement (AMS Police; Sapir, 2005, point 270, p. 78-79).

18. *Details:* This criterion refers to the usage of details in the statement. If the writer can produce a lot of details in the statement, the statement is expected to be truthful (Bockstaele, 2008a, 2008b).
19. *Unasked explanation:* This criterion refers to an explanation why something happened, given by the writer, without asking. According to SCAN this information is very sensitive (AMS Police; Sapir, 2005, point 299a, p. 86-87).
20. *Lack of conviction or memory:* This criterion refers to vagueness in the statement about certain event (e.g. "I think…", "I suppose…") or parts in the statement where the writer states that he or she cannot remember something. These cues are seen as an indication of deceit (Vrij, 2008b).
21. *Unexpected complication:* This criterion refers to the presence of unexpected events. For example, a rape victim who states that her attacker first had to get the cat of the bed. It is expected that the presence of this criterion indicates truthfulness (Bockstaele, 2008a, 2008b).
22. *Denial of allegation*: When the writer directly denies the allegation in the statement it indicates truthfulness (e.g. "I did not…"). For deceptive persons, it is expected they will not directly deny the allegation (e.g. "Do you think I would do something like that?") (Vrij, 2008b; Sapir, 2005, point 206, p. 57).
23. *Spontaneous corrections:* This criterion refers to the presence of corrections in the statement. Before examinees writer their statement they are instructed not to cross anything out. When examinees do not follow this instruction, it may indicate deception (Vrij, 2008b).
24. *Sensory perceptions:* The use of sensory perceptions within a statement is an indication of truthfulness (Bockstaele, 2008a, 2008b).
25. *Unresponsiveness to questions*: When a writer refuses to write down information this is an indication of deception (Bockstaele, 2008b).
26. *Avoiding answers:* The unwillingness of examinees to give direct answers to questions indicates deception (Bockstaele, 2008b).
27. *Activities:* According to SCAN certain discussed activities are important. These activities include brushing teeth, turning the light on or off, closing or opening a door or getting in or out a car. These activities can give information about deception or child sexual abuse (AMS Police; Sapir, 2005, point 303, p. 89).
28. *Exact location:* When a writer gives an exact location of another person in the statement this gives an indication about a conflict between the writer and the other person (AMS Police).
29. *Negative language use:* When a writer gives information about something that did not happen, thus when a sentence is presented in negative. This is sensitive information for the writer. (AMS Police; Sapir, 2005, point 299c, p. 87). This criterion is a combination of "Denial of allegation" and "Lack of conviction or memory".

4

Scientific Content Analysis (SCAN) cannot distinguish between truthful and fabricated accounts of a negative event

ABSTRACT

The Scientific Content Analysis (SCAN) is a verbal veracity assessment method currently used by investigative authorities worldwide. Yet, research investigating the accuracy of SCAN is scarce. The current study tested whether SCAN could accurately discriminate between true and fabricated statements. To this end, a total of 117 participants were asked to write down one true and one fabricated statement about a recent negative event that happened in their lives. All statements were analysed using 11 criteria derived from SCAN. Results indicated that SCAN was not able to correctly classify true and fabricated statements. Lacking empirical support, the application of SCAN in its current form should be discouraged.

Published as:
Bogaard, G., Meijer, E., Vrij, A., Broers, N. J., & Merckelbach, H. (2016). Scientific Content Analysis (SCAN) cannot distinguish between truthful and fabricated accounts of a negative event. *Frontiers in Psychology*, *7*, 243. doi: 10.3389/fpsyg.2016.00243

4.1 INTRODUCTION

Research has revealed that nonverbal cues (e.g., behavioral cues such as gaze aversion, sweating) are faint and differences between liars and truth tellers are small at best (DePaulo et al., 2003; Sporer & Schwandt, 2007). However, findings about verbal cues are less variable and are more strongly related to deception (C. F. Bond & DePaulo, 2006; Vrij, 2008b, 2008c). Verbal cues (or content cues) are cues that can be found in the content and meaning of a statement, such as the number of details that are included in a story (e.g., he had a large spider tattoo in his neck). Indeed, lying has been shown to result in qualitative differences between deceptive and truthful language. As a result, various verbal credibility assessment tools have been developed that address these content criteria within statements. Although the exact content criteria included may differ depending on the method, the procedure is highly similar. The presence of the criteria within the statements is carefully checked, and based on the presence of absence of the various criteria, a conclusion is drawn about its truthfulness.

One example of such a content criterion is "quantity of details". To fulfil this criterion, a statement has to be rich in details, such as mentioning places (e.g., it happened in the kitchen), times (e.g., on Sunday evening at 8 p.m.), descriptions of people and objects (e.g., a tall man with bright blue eyes), etc. Additionally, deceit has been related to the use of fewer personal pronouns (e.g., using "the house" instead of "our house") and fewer negations (e.g., no, never, not), using less perceptual information (e.g., "I could smell the alcohol in his breath"), less details overall and shorter statements (Amado et al., 2015; Hauch, Blandón-Gitlin, Masip, & Sporer, 2014; Masip et al., 2005; Newman et al., 2003). As mentioned previously, several methods have been developed to address these issues.

Two well-established credibility assessment tools that tap into such language differences are the Criteria Based Content Analysis (CBCA) and Reality Monitoring (RM). For CBCA, two theoretical assumptions have been presented by Köhnken (1996). First, lying is seen as more cognitively challenging that telling the truth. Secondly, liars are expected to be more concerned with impression management than truth tellers. More precisely, a first subset of CBCA criteria is included because they are deemed too difficult to fabricate (e.g., descriptions of interactions with the perpetrator). Hence, their presence in a statement indicates an actual experience. The remainder of the CBCA criteria are concerned with the way an interviewee presents his or her story. It is expected that liars are concerned with how they are viewed by others and therefore leave out information that can possibly damage their view of being an honest person (e.g., mentioning self-deprecating information). Consequently, a truthful person is more likely to include these criteria in their statement than a deceptive person. RM, in contrast, is derived from memory research and holds that memories of real events are obtained through sensory processes, making them more clear, sharp and vivid, while fabricated statements are the result of fantasy and are usually vaguer and less concrete (Johnson

& Raye, 1981). Indeed, various studies reported supportive evidence for these methods. Their overall accuracy for detecting deceit varies around 70%, and is considerably higher than chance level (Amado et al., 2015; Johnson & Raye, 1981; Masip et al., 2005; Steller & Köhnken, 1989; Undeutsch, 1967; Vrij, 2005).

Despite the research showing above chance accuracy for CBCA and RM, their field use seems limited. A third method - that is employed by Law enforcement worldwide - is Scientific Content Analysis (SCAN). SCAN was developed by former Israeli polygraph examiner Avinoam Sapir (2005), who – based on his experience with polygraph examinees - argued that people who tell the truth differ from liars in the type of language they use. Based on these assumed differences, Sapir developed criteria that, according to him, can assist in differentiating between true and fabricated statements, but without reporting a theoretical foundation as to why these specific criteria should differ. For example, SCAN includes the criterion "social introduction". It is argued that people who are described in a statement should be introduced with name and role (e.g., My friend, John). If a person leaves out information (e.g., We stole the key), so leaving out the name, role or both, this indicates deception. Another criterion is the "structure of the statement". According to SCAN, 20% of the statement should consist of information that led up to the event, 50% should be about the main event and 30% of the statement should be about what happened after the event. The more the statement deviates from this structure, the higher the likelihood that the statement is deceptive. In contrast to CBCA and RM, no theoretical rationale is presented, and there is no evidence that these criteria are actually diagnostic (Bogaard, Meijer, & Vrij, 2014; Nahari et al., 2012; Vanderhallen et al., 2015).

Research about SCAN is scarce, although the method is applied worldwide (e.g., Australia, Belgium, Canada, Israel, Mexico, UK, US, the Netherlands, Qatar, Singapore, South Africa) and is also used by federal agencies, military law enforcement, private corporations, and social services (retrieved from www.lsiscan.com/id29.htm). Moreover, the third author asked during an investigative interviewing seminar which lie detection tool was used by the practitioners in the audience. These practitioners came from many different countries and the most frequent answer was SCAN (Vrij, 2008b). In a typical SCAN procedure, the examinee is asked to write down "everything that happened" in a particular period of time, to get a "pure version" of the facts (Sapir, 2005). This pure version is typically obtained without the interviewer interrupting or influencing the examinee. Next, a SCAN trained analyst investigates a copy of the handwritten statement, using several criteria that are described throughout the SCAN manual (Sapir, 2005). Criteria that are present within the written statements are highlighted according to a specific colour scheme, circled or underlined. The presence of a specific criterion can either indicate truthfulness or deception, depending on the criterion itself. This SCAN analysis serves as a basis to generate questions that could elucidate important details within the statement, and/or to make a judgment of the veracity of the statement. Although SCAN is used worldwide, it lacks a well-defined list of criteria, as well as

a standardized scoring system. Bogaard et al. (2014b) showed that 12 criteria primarily drove SCAN in sexual abuse cases, largely overlapping with the criteria list described in Vrij (2008b) (Chapter 3). Only six published studies examined the validity of SCAN (Bogaard, Meijer, & Vrij, 2014; Driscoll, 1994; Nahari et al., 2012; Porter & Yuille, 1996; Smith, 2001; Vanderhallen et al., 2015) of which only four were published in peer reviewed journals. The two studies that were not published in peer reviewed journals [Driscoll (1994) and Smith (2001)] were both field studies investigating actual statements of suspects.

Driscoll (1994) investigated 30 statements that were classified as either apparently accurate or doubtful. With the help of SCAN, 84% of the statements could be correctly classified. In the study of Smith, five groups of experts were asked to analyse 27 statements. These statements were previously classified by police officers as truthful, false or undecided. This was done on the basis of confessions and supporting evidence. Three groups consisted of SCAN trained officers that had minimal, moderate or extensive experience with using SCAN. The two other groups consisted of newly recruited officers and experienced officers. The first three groups used SCAN to analyse the statements, while the latter two groups judged the veracity of the statements without using SCAN. Overall, the SCAN groups correctly judged 78% of the statements, which was similar to the accuracy of the experienced officers. At first glance, these results seem to support SCAN. Yet, in both studies ground truth of the statements was unknown; statements were categorized as either truthful or doubtful without having hard evidence supporting this categorisation. Moreover, it cannot be excluded that the SCAN outcome influenced the course of the investigation, and therefore the confessions and supporting evidence that were gathered. A typical problem that can occur in such studies is that errors are systematically excluded from the sample. For example, if a statement is erroneously judged as truthful, no further investigation takes place. This means no evidence will be found revealing that an error has been made, and such erroneous classifications are then excluded from the sample. This way of selecting the sample may therefore be biased to overestimate SCAN's accuracy (for more information see Iacono, 1991; Meijer, Verschuere, Gamer, Merckelbach, & Ben-Shakhar, 2016). Moreover, in Smith's study, it was unclear whether the three undecided statements were included in the reported analyses (Armistead, 2011).

Four studies investigating SCAN were published in peer-reviewed journals. Porter and Yuille (1996) addressed the problem of ground truth by asking participants to commit a mock crime. However, they only investigated three SCAN criteria (i.e., unnecessary connectors, use of pronouns and structure of the statement), and results indicated no significant differences between true and fabricated statements concerning these criteria. Nahari et al. (2012) asked six independent raters to assess the presence of 13 SCAN criteria within various true and fabricated statements. Yet, SCAN did not discriminate between truthful and fabricated statements, a conclusion that was also supported by Bogaard et al. (2014). In their study, participants were asked to write

down one truthful and one fabricated autobiographical statement about a negative event that recently happened to them. Two raters indicated the presence of 12 SCAN criteria, but no significant differences emerged between truth tellers and liars. Vanderhallen et al. (2015), finally, asked SCAN trained police officers to classify four statements as either truthful or deceptive based on SCAN, and compared their accuracy to students and police officers who made this classification without the help of SCAN. The SCAN group had an average accuracy of 68%, police officers without SCAN 72%, and students 65%. The accuracy of the SCAN group did not significantly differ from the police officers who did not use SCAN. The authors concluded that SCAN did not have an incremental value in detecting deceit.

Given that SCAN is used worldwide in police investigations, providing support, or the lack thereof, is not trivial (Meijer et al., 2009). Using a data set of 234 statements, the current study aimed at extending previous SCAN findings, and to investigate whether the different SCAN criteria can discriminate between truthful and fabricated statements. Although Nahari et al. (2012), Bogaard et al. (2014) and Vanderhallen et al. (2015) investigated SCAN, Bogaard et al. mainly focused on the SCAN total scores, and not on the separate criteria, or the accuracy of SCAN. Separate criteria scores were reported, but their power was too low to draw any conclusions from these results. In contrast, Nahari et al. asked participants to perform a mock crime, meaning that the statements that were analysed with SCAN were restricted to "false denials" (i.e., people who performed the mock crime but lied about it). Moreover, in the study of Vanderhallen et al. four statements on traffic accidents were used. The statements included in our study are broader than false denials or traffic accidents, as we requested participants to write about a negative autobiographical event. In this way, participants not only reported false denials, but also false allegations (i.e., stating they fell victim to a crime, while in fact they were not). Participants could report about whatever they preferred, thereby including various topics, as would also be the case in police investigations where SCAN would be applied.

4.2 METHOD

Participants

All participants (*n* =117) were first and second year health sciences students (i.e., Mental Health or Psychology) of Maastricht University (37 men). The data of 85 participants were collected specifically for this study, while the remaining 32 came from the control group of Bogaard et al. (2014; Chapter 6). Instructions for these two datasets were identical, and they were combined to increase power. We report the analysis for the entire sample below, but also include the finding for the new dataset in Appendix E.

Participants could choose whether they wanted to receive one course credit or a 7,5 Euro gift voucher for their participation. Approximately 50 students chose the gift voucher over the course credit. All participants read and signed a letter of Informed Consent before they took part in this study. Participants had a mean age of 21 years (*SD* = 2.35). The experiment was approved by the standing ethical committee.

Procedure

Upon arrival in the lab, participants were told that the study was about the accuracy of verbal lie detection methods. Participants were asked to write about a truthful and a fabricated event. The order in which participants wrote these statements was randomized. Approximately half of the participants started with the truthful statement, the other half started with the fabricated story. For the truthful statement participants received the following instruction: "For this study we ask you to think about an event you actually experienced. More specifically, this event should be about a recent negative experience; think about a financial, emotional or physical negative event you've been through the past months." For the fabricated statement participants received the following instruction: "For this study we ask you to think about a negative event that you have not actually experienced. This event should be about a recent negative experience; think about a financial, emotional or physical negative event you could have been through the past months. This event should not be based on something that happened to you or your friends or family. Please pretend as if this event took place somewhere in the previous months. Although the story should be fabricated, the statement should consist of a realistic scenario." After the instruction, participants had the opportunity to think about a real and a fabricated story for a maximum of 5 minutes. Participants were assured that their stories would be treated confidentially and anonymously. They were told that the length of the stories should be approximately one written page (A4). No time limit was set to produce the statements.

Statement coding

After participants finished their stories, these were analysed by four raters. One rater completed the three-day SCAN course. The other three raters received a 2-hour training about SCAN, using the SCAN manual (Sapir, 2005), given by the SCAN trained rater. Moreover, they received the appropriate pages (Chapter 10; 282-287) of Vrij (2008a) about SCAN. During the training, all 12 criteria were discussed separately and examples of the specific criteria were presented and discussed. Next, raters received two practice statements of 1 page each, and were asked to analyse these statements. After all raters analysed these statements, their analyses were discussed and questions they still had about SCAN were answered. When the training was completed, raters started analysing the statements.

Although the raters were not blind to the aim of the study, they were blind to the veracity of the statements. The first author served as one of the raters, the other raters were not otherwise involved in the study and were research assistants of the first author. The rater who completed the original SCAN training scored all 234 statements, while the other three raters scored approximately 80 statements each. To control for potential order effects, the sequence of the statements to be scored was varied from rater to rater. Rater A scored all statements in the order of 1 to 234, while the other raters scored the statements in the reverse order (rater B started from 79 to 1, rater C started from 157 to 80 and rater D from 234 to 158).

A total of 12 criteria (see Chapter 3) were coded within the statements. According to SCAN, seven of these criteria indicate truthfulness: (1) Denial of allegations, (2) Social introductions, (3) Structure of the statement, (4) Emotions, (5) Objective and subjective time, (6) First person singular, past tense, (7) Pronouns, while the remaining five indicate deception: (8) Change in language (9) Spontaneous corrections (10) Lack of conviction or memory (11) Out of sequence and extraneous, (12) Missing information. See Appendix D for a complete description of the different criteria. All criteria that are expected to indicate truthfulness were scored on a 3-point scale ranging from 0 (not present) to 2 (strongly present), while the 5 criteria expected to indicate deception were scored in reverse, ranging from -2 (strongly present) to 0 (not present). With this scoring system, a higher score indicates a higher likelihood that the statement is truthful and vice versa. By providing participants with a fixed list of SCAN criteria, their definitions, and a fixed coding scheme, we aimed to resolve the low inter-rater reliability reported in Chapter 3. This is important, because only when inter-rater reliability is sufficient, we can investigate SCAN's ability to discriminate between truthful and fabricated statements.

4.3 RESULTS

Inter-rater reliability

Inter-rater reliability was calculated by means of Cohen's Kappa for each of the 12 separate criteria. The Kappa values for the truthful statements varied from 0.60 to 1 with an average Cohen's Kappa of 0.77. The Kappa values for the fabricated statements varied from 0.65 to 1, with an average kappa of 0.78. These results indicated that there is high agreement between the raters (Landis & Koch, 1977). Because variance was low for several criteria, Cohen's Kappa could give a distorted image of the actual inter-rater reliability. Therefore, we also included inter-rater agreement calculated by means of percentage agreement and its presence in the statement. To this end, we dichotomized the original data set with presence coded as 1 and absence as 0. High agreement was achieved for all SCAN criteria ranging from 80% to 100% with an average of 90%. The

scoring of the three raters was always compared to those of the rater that completed the SCAN training. As reliability proved to be sufficient, this also indicated that our 2-hour SCAN training was sufficient to score the investigated SCAN criteria reliably.

Data analysis

Because the inter-rater reliability was high, we averaged the scores of the two raters for each criterion. Due to the nature of our instructions (i.e., autobiographical statements) the first criteria could not be coded in the statements. Thus, we have left out "denial of allegations" in the following analysis. Next, we calculated the sum scores for each statement by summing up the averaged scores of the separate criteria. To investigate whether SCAN can accurately discriminate between truthful and fabricated statements, we conducted several Generalized Estimation Equation (GEE) analyses (see for example Burton, Gurrin, & Sly, 1998); one for each separate criterion. Moreover, we conducted a paired samples *t*-test for the sum score, and a discriminant analysis to test SCAN's predictive power concerning the veracity of statements.

Number of words

The length of the statements did not differ significantly between the true (M = 265.42; SD = 85.48) and fabricated statements (M = 261.86; SD = 88.12) [$t(116) = 0.63$, $p = 0.53$, $d = 0.04$].

Table 4.1. Means, standard deviations and percentage present for each SCAN criterion as a function of veracity.

SCAN criteria	True			Fabricated		
	Mean	SD	% present	Mean	SD	% present
Denial of allegations	0	0	0	0	0	0
Social introduction	1.26	.81	76.90	1.40	.71	87.20
Spontaneous corrections	-.61	.62	56.40	-.64	.63	58.10
Lack of conviction or memory	-.16	.36	18.80	-.14	.33	16.40
Structure of the statement	.73	.60	67.50	.59	.60	56.40
Emotions	1.05	.62	83.80	.95	.65	76.10
Objective and Subjective time	.71	.65	62.40	.79	.65	69.20
Out of sequence and extraneous info	-.18	.38	21.40	-.20	.43	22.20
Missing information	-.64	.55	75.00	-.67	.52	67.50
First pers sing, past tense	1.59	.63	92.30	1.60	.60	94.00
Pronouns	1.68	.49	97.40	1.69	.50	97.40
Change in language	**-.09**	**.27**	**12.00**	**-.23**	**.43**	**24.80**

Note. Significant difference between statement types, $p = 0.01$ is in bold.

SCAN criteria scores

Table 4.1 shows the mean differences in each of the SCAN criteria as a function of veracity. Next, we analysed the data with GEE to investigate the differences between truthful and fabricated statements for each separate criterion. In doing so, we had to dichotomize our data by recoding presence as 1 (regardless of whether the score was a 1 or a 2) and absence as 0. Due to very low variability of the criterion "pronouns" (i.e., it was present in almost all of the statements), this criterion was left out of the analysis. To correct for multiple testing, we used an alpha level of .01. As Table 4.2 shows, only one criterion significantly differed between the statements, namely "Change in language". Participants included more changes in language in their fabricated statements compared with their truthful statements. This criterion was present in 29 out of 117 fabricated statements (24.8%) and in 14 out of 117 true statements (12%). In Appendix E (Table E1) we have presented the results of only the new data, and again only "Change in language" significantly differed between statements.

Table 4.2. Overview of parameters from the GEE analysis.

Criteria	Beta Estimate	SE	95% CI	Odds ratio
Change in language	-0.89	0.36	-1.59, -.18	**0.79**
Social introduction	-0.71	0.34	-1.37, -0.05	0.51
Emotions	0.48	0.27	-0.06, 1.03	0.23
Structure of statement	0.47	0.26	-0.04, 0.99	0.22
Objective and subjective time	-0.31	0.21	-0.73, 0.12	0.10
First pers. sing., past tense	-0.27	0.51	-1.26, 0.72	0.07
Lack of conviction or memory	0.24	0.31	-0.36, 0.85	0.06
Missing information	-0.12	0.21	-0.53, 0.30	0.01
Spontaneous corrections	-0.07	0.22	-0.50, 0.36	0.00
Out of sequence and extraneous information	-0.05	0.28	-0.60, 0.50	0.00

Note. Significant difference between statement types, $p = 0.01$ is in bold.

SCAN sum scores

There were no differences in SCAN sum scores between true ($M = 5.33$; $SD = 2.10$) and fabricated ($M = 5.15$; $SD = 2.25$) statements [$t(116) = 0.77$, $p = 0.44$, $d = 0.12$]. Moreover, we also conducted a discriminant analysis to investigate whether the SCAN criteria could predict veracity. As can be seen in Table 4.3, only one significant mean difference was observed for "Change in language" ($p < 0.01$). The discriminant function revealed a low association between veracity and SCAN criteria, only accounting for 7.20 % of the variability. Closer analysis of the structure matrix revealed that three criteria had moderate discriminant loadings (i.e., Pearson coefficients): these were – again - "Change in

language" (0.66), "Structure of the statement" (0.41), and "Social introduction" (-0.35). The uncorrected model resulted in correct classification of 59% of the truth tellers, and 65% of the liars. The cross-validated classification, however, showed that 49.60 % of the liars and 53 % of the truth tellers were correctly classified, thereby showing that SCAN performed around chance level. In Appendix E (Table E2), we have presented the results for only the new data, and results were similar. The uncorrected model resulted in correct classification of 63% of the truth tellers, and 58% of the liars. The cross-validated classification showed that 50 % of the liars and 55 % of the truth tellers were correctly classified, again showing that SCAN performed around chance level.

Table 4.3. Detailed overview of discriminant analysis coefficients.

Criteria	Mean	SD	Structure matrix	Discriminant function coefficients
Change in language	-0.16	0.37	0.66	1.82
Structure of statement	0.66	0.60	0.41	0.79
Emotions	1.00	0.63	0.29	0.67
Spontaneous corrections	-0.62	0.62	0.07	0.23
Out of sequence and extraneous information	-0.19	0.40	0.10	0.20
Missing information	-0.65	0.53	0.12	0.12
Pronouns	1.69	0.49	-0.03	-0.09
First pers sing. past tense	1.59	0.61	-0.04	-0.14
Lack of conviction or memory	-0.15	0.35	-0.11	-0.21
Objective and subjective time	0.75	0.65	-0.23	-0.43
Social introduction	1.33	0.77	-0.35	-0.53

4.4 DISCUSSION

In the current study, we failed to find support for SCAN as a lie detection method. The total SCAN score did not significantly differ between true and fabricated statements, confirming previous results (Bogaard et al., 2014; Nahari et al., 2012). Interestingly, for a subset of our data CBCA and RM sum scores were coded and these did discriminate between the truthful and fabricated statements (Bogaard et al., 2014; chapter 6). It seems that the absence of significant SCAN findings cannot be attributed to the quality of the statements used in this study. Furthermore, we investigated the separate SCAN criteria, and only one criterion "Change in language", differentiated significantly between true and fabricated statements; participants changed their language more in their fabricated statements compared with their truthful statements.

The criterion 'change in language' is not described in other verbal credibility methods (e.g., CBCA, RM). Therefore, our findings concerning this criterion are noteworthy.

Sapir (2005) explained in his manual that especially words describing family members (e.g., mother, father, dad, mom, etc.), people (e.g., someone, individual, man, guy, etc.), communication (e.g., told, spoke, talked, etc.), transport (e.g., vehicle, car, truck, etc.), and weapons (e.g., gun, rifle, revolver, pistol, etc.) should be investigated carefully. The idea is that a change with regard to how these elements are described indicates something has altered in the mind of the writer. When the events in the statements justify this change, it does not indicate deception per se. However, in all other cases these changes indicate deceit. What exactly is meant by a justification is not described in the manual. Consequently, due to the absence of clear guidelines on verifying whether a change is justified, the current study scored all changes in language as a cue to deceit, and might therefore differ from how SCAN is used in practice.

Both the analyses of the SCAN sum score and the discriminant analysis showed SCAN did not perform above chance level. This chance level performance can be understood when looking at various contradicting interpretations of its criteria compared with CBCA. More precisely, both methods describe 'spontaneous corrections' and 'lack of conviction or memory', but differ in their use. CBCA interprets both criteria as a sign of truthfulness, while SCAN interprets both criteria as a sign of deceit. Logically, only one interpretation can be correct. As CBCA is far more embedded in the scientific literature and has been shown to detect deceit above chance level (Amado et al., 2015; Vrij, 2005), CBCA's interpretations should be favoured over SCAN. Also, SCAN does not consider criteria involved in judging distinctive types of details. Both CBCA and RM consist of various types of details that have to be checked. For example, with these methods it is checked whether there is information in the statement about when (i.e., temporal details) and where (i.e., spatial details) the event took place, about what the writer saw during the event (i.e., visual details), and whether there were any other perceptual details (i.e., smells, tastes, sensations, sounds). Research showed that especially these types of criteria are significantly more present in truthful compared to fabricated statements (DePaulo et al., 2003; Masip, Sporer, et al., 2005; Vrij, 2005).

Relatedly, recent meta-analytical research reveals that passively observing cues contributes only in a limited way to our deception detection abilities, as most of these cues are generally weak (Hartwig & Bond, 2011). Hartwig and Bond (2011) argue that we should actively increase the verbal and nonverbal differences between liars and truth tellers. Various techniques have already been suggested, such as focusing on unanticipated questions during the interrogation (Vrij et al., 2009), applying the Strategic Use of Evidence technique (Granhag, Strömwall, & Hartwig, 2007) or inducing cognitive load (Vrij, Fisher, Mann, & Leal, 2006, 2008; Vrij, Granhag, Mann, & Leal, 2011; Vrij, Leal, Mann, & Fisher, 2012). SCAN fails to actively influence the information that is provided by the interviewee, which potentially contributes to its chance performance.

Finally, advocates of SCAN may argue that the way SCAN is tested in laboratory studies such as these, is far from how it is applied in the field, and that the results therefore do not translate. However, the diagnostic value of SCAN and its criteria lies within its

capabilities of discriminating between truthful and fabricated statements. SCAN makes no assumptions as to *why* or *when* these differences between truths and lies occur, only that they occur. As such, laboratory studies – for example where participants are asked to fabricate a negative event – should also be able to pick up such differences, if they arise at all. Moreover, it has proven to be exceptionally difficult to test the accuracy of SCAN in field studies as the reliability of SCAN has shown to be extremely low (Bogaard et al., 2014b; Chapter 3 ; Vanderhallen et al., 2015). The only way to control for this low reliability is to use a more standardized scoring system, as we have done in the current study. For example, as is mentioned previously, SCAN does not consist of a fixed list of criteria, and the criteria are not scored on a scale. In field studies, SCAN analysts write a report about the presence or absence of the criteria and on the basis of this report, they draw a conclusion about the truthfulness of the statement. It is unclear how many criteria are taken into consideration when making a judgment, and whether these criteria are weighed equally.

In sum, SCAN has no empirical support to date, and fails to include criteria investigating different types of details. Only one criterion showed potential for lie detection, but should be investigated more thoroughly to overcome the problems that are inherent to SCAN and its criteria (e.g., vague description, ambiguous interpretation). As a result, we discourage the application of SCAN in its current form.

Appendix D

SCAN criteria (derived from Vrij (2008a)).

1. *Denial of allegations*: Refers to whether the examinee directly denies the allegation in the statement by stating "I did not…". This criterion assumes that a truthful person is more likely to directly deny his or her involvement in the act.
2. *Social introduction*: Refers to how the persons described in the statement are introduced. People that are described within a statement should be introduced in an unambiguous way, usually by mentioning their name and role (e.g., My wife, Susan). Deviations from this type of introduction indicate deception.
3. *Spontaneous corrections*: Refers to all corrections that are made within the statements. Before the writer start with the statement s/he is instructed not to cross anything out, when the writer fails to follow this instruction, this indicates deception.*
4. *Lack of conviction or memory*: Refers to when the writer is vague about certain elements within the statement (e.g. "I think…", "I guess…") or when the writer admits he or she has forgotten something (e.g., "I do not remember how we got to the house"). Lack of memory indicates deceit.*
5. *Structure of the statement*: Refers to the balance of the statement. In a truthful statement 20% is used to describe activities leading up to the event, the next 50% to describe the actual event, and the final 30% to discuss what happened after the event.
6. *Emotions*: Refers to where there are emotions described in the statement. Usually emotions should be described in the epilogue of the statement. When emotions are already included within the description of the prologue (before the actual event), this indicates deception. For example, "On Saturday something strange happened to me, I was really scared" (emotions before main event) or "when he was gone, I felt disgusted with myself" (emotions after the main event). The former would indicate deception, the latter truthfulness.
7. *Objective and subjective time*: Refers to how different time periods are covered in the statement. Objective time refers to the actual duration of events described, whereas subjective time refers to the number of words spent to describe these events. On average a writer is expected to need three or four lines per hour when describing one day. Large deviations from this pace suggest deception.
8. *Out of sequence and extraneous information*: Examines whether the statement includes information that is given by the writer but has no apparent meaning for the reader or whether there is strange or irrelevant information within the statement. Whether the information is seen as strange or irrelevant depends on the statement itself. It is thought that by including this type of information, the writer is distracting the reader to hide more important information. This is seen as a sign of deception.*

9. *Missing information*: Refers to phrases in the statement that indicate that some information has been left out. For example, words such as "after a while", "shortly thereafter", or "the next thing I remember" all indicate there is information missing within the statement. This is especially relevant when the writer is discussing the main event. Missing information during the main event could indicate that the writer is deliberately hiding information, which indicates the person is deceptive.*
10. *First person singular, past tense*: Refers to the format in which a statement is written. This is called the test of commitment, and holds that a truthful person will write the statement in first person singular, past tense. Deviations from past tense or writing in the third person could indicate a lack of commitment and hence could indicate deception.
11. *Pronouns*: Refers to the use of pronouns in the statement. When pronouns (e.g., "he", "mine", "my") are missing in the statement, or more pronouns are expected, this could suggest that the writer wants to distance him/herself from the statement. This indicates deception. For example, when a writer refers to his car as "the car" and never as "my car" this could mean he is being deceptive about what happened to the car.
12. *Change in language*: Refers to the change of terminology or vocabulary in the statement. This is especially important for words that are related to categories such as family members, people, communication, transport or weapons. When a change of language is obvious in a statement (e.g., knife to blade) but no justification is given for such a change, this indicates deception. A change in language indicates that something has altered in the mind of the writer.*

Appendix E

Table E1. Overview of parameters from the GEE analysis of new data.

Criteria	Beta Estimate	SE	95% CI	Odds ratio
Change in language	-1.08	0.44	-1.94, -.23	**1.18**
Social introduction	-0.87	0.52	-1.37, -0.05	0.75
Emotions	0.73	0.43	-0.12, 1.57	0.53
Structure of statement	0.45	0.31	-0.16, 1.07	0.21
Objective and subjective time	-0.46	0.28	-1.01, 0.10	0.21
First pers sing. past tense	-0.31	0.54	-1.38, 0.75	0.10
Lack of conviction or memory	0.20	0.33	-0.46, 0.86	0.04
Missing information	-0.21	0.33	-0.87, 0.43	0.05
Spontaneous corrections	-0.21	0.29	-0.78, 0.35	0.04
Out of sequence and extraneous information	-0.08	0.36	-0.80, 0.63	0.01

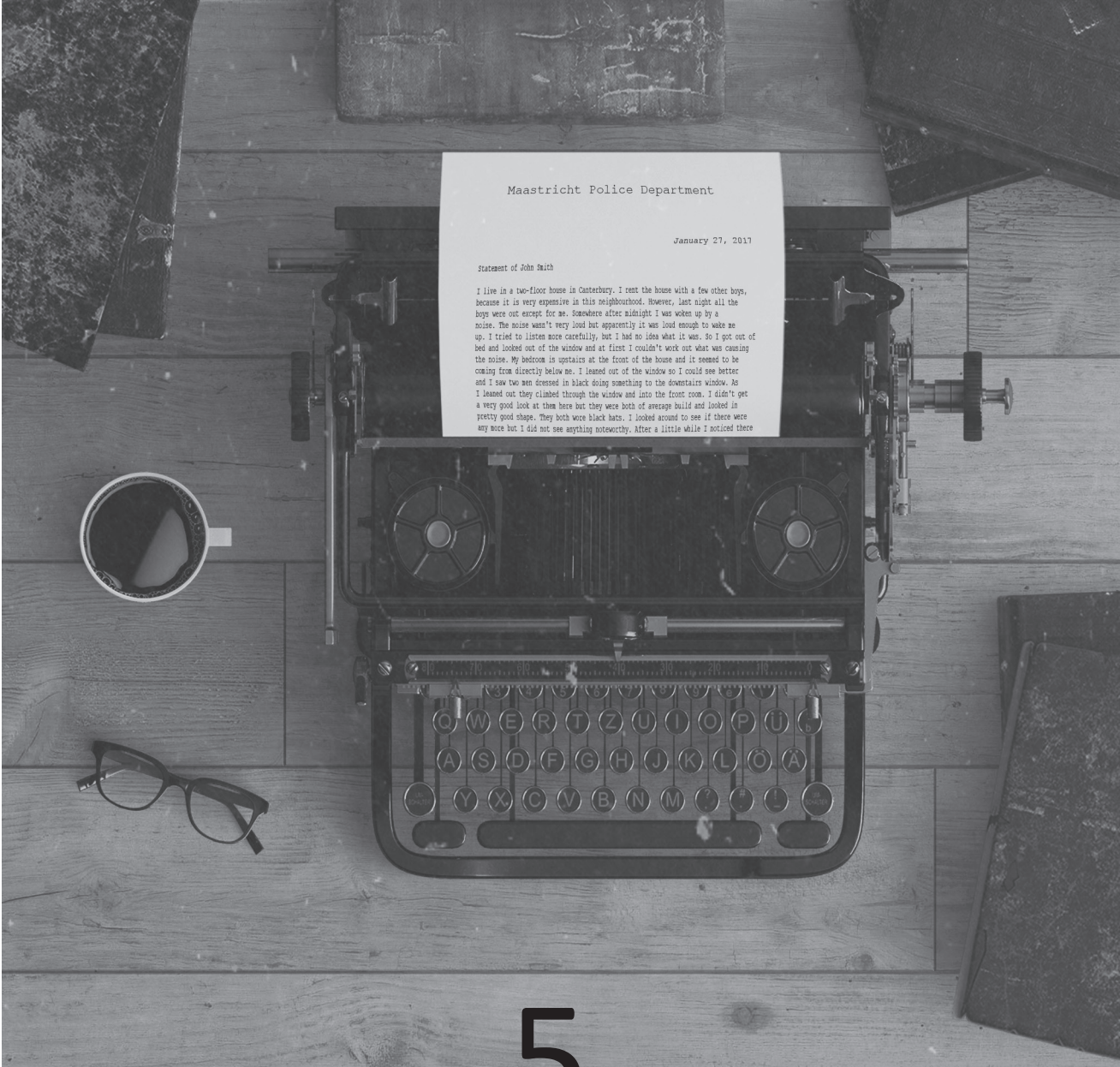
Note. Significant difference between statement types, $p = 0.01$ is in bold.

Table E2. Detailed overview of discriminant analysis coefficients derived from the new data.

Criteria	Mean	SD	Structure matrix	Discriminant function coefficients
Change in language	-0.17	0.38	0.05	1.51
Structure of statement	0.64	0.60	0.32	0.80
Emotions	1.11	0.60	-0.24	0.74
Spontaneous corrections	-0.71	0.63	0.36	0.34
Out of sequence and extraneous information	-0.16	0.41	0.14	0.48
Missing information	-0.80	0.50	0.12	0.23
Pronouns	1.69	0.51	-0.07	0.22
First pers sing. past tense	1.59	0.63	0.10	-0.42
Lack of conviction or memory	-0.20	0.39	-0.34	0.13
Objective and subjective time	0.91	0.66	-0.19	-0.66
Social introduction	1.56	0.69	0.56	-0.66

Part 2

SCAN and alternative credibility assessment methods



5

Contextual bias in verbal credibility assessment: CBCA, RM and SCAN

ABSTRACT

Verbal credibility assessment encompasses several methods used to evaluate the credibility of statements by examining their content. In two experiments, we tested to what extent these methods are sensitive to contextual bias. Four statements were presented, while their context was manipulated by confronting raters with extra-domain information that either enhanced or diminished the credibility of the statements. In Experiment 1, 32 police officers analysed the statements using Scientific Content Analysis. In Experiment 2, 128 undergraduates analysed the statements using criteria derived from Criteria Based Content Analysis, Reality Monitoring or Scientific Content Analysis. Results showed that all three methods were equally vulnerable to contextual bias.

Published as:
Bogaard, G., Meijer, E., Vrij, A., Broers, N. J., & Merckelbach, H. (2013). Contextual Bias in Verbal Credibility Assessment: Criteria Based Content Analysis (CBCA), Reality Monitoring (RM), and Scientific Content Analysis (SCAN). *Applied Cognitive Psychology*, *28*, 79–90. doi: 10.1002/acp.2959

5.1 INTRODUCTION

People are not very successful in detecting lies. An extensive body of research shows that when they base their judgements on verbal and nonverbal behavior, individuals, including trained police officers, generally perform only just above chance level (Aamodt & Custer, 2006; C. F. Bond & DePaulo, 2006; Elaad, 2003; Vrij, 2008b). Nonetheless, judging the veracity of statements of victims, suspects, and witnesses plays an important role in the criminal justice system. To facilitate the detection of deceit in such statements, several methods of credibility assessment based on verbal indicators have been developed.

These methods aim to discriminate between true and false statements not by looking at their source (i.e., the person issuing the statement) but rather by focussing on the language qualities of the statements. One such method is the Scientific Content Analysis (SCAN; Sapir, 2005). SCAN was developed by former Israeli polygraph examiner Avinoam Sapir (2005), who argued that truth tellers and liars differ in the type of language they use. Based on these alleged differences, Sapir developed a list of criteria that could assist in differentiating between true and false statements. Most SCAN criteria are thought to be more present in false than in true statements.

SCAN is the most frequently used verbal credibility assessment method worldwide (Vrij, 2008b). Four studies examined SCAN, but found no solid evidence for its discriminative value (Driscoll, 1994; Nahari, Vrij, & Fisher, 2012; Porter & Yuille, 1996; Smith, 2001). In addition, SCAN has low inter-rater reliability (Smith, 2001), which means that users differ in the way they apply SCAN. The list of SCAN criteria is extensive, and no standardised set exists yet (Bogaard et al., 2014b). In addition, different users employ different criteria when assessing the same statement (Smith, 2001). The unstandardized nature of SCAN raises the suspicion that it may be sensitive to contextual or expectancy bias (e.g., Risinger, Saks, Thompson, & Rosenthal, 2002)

Contextual or expectancy bias refers to a set of phenomena that all have in common that when experts are exposed to contextual information, it may shift their decision thresholds as a function of the expectations that they implicitly generate on the basis of the context information (Risinger et al., 2002). One straightforward example is confirmation bias, which is the tendency to search for evidence that confirms an a-priori held belief, while ignoring evidence that disconfirms it (Findley & Scott, 2006; Jones & Sugden, 2001). In addition to searching for confirming evidence, confirmation bias also includes the tendency to judge information supporting one's beliefs as more important than disconfirming information (Findley & Scott, 2006). In sum, it refers to an implicit selectivity in the acquisition and usage of evidence (Nickerson, 1998).

The negative consequences of contextual bias effects have been well documented in the forensic domain with diagnostic methods that have a longer track record than SCAN. Findley and Scott (2006), for example, give an extensive overview of how such biases play a crucial role in miscarriages of justice. As an illustration, Dror, Charlton, and

Péron (2006) investigated the effect of supplying fingerprint experts with misleading information about the context of the fingerprint they had to evaluate. Participants were asked to examine a pair of fingerprints that they had judged five years earlier as a clear "match". However, the prints were now presented in a context that suggested a non-match. Supplying this false information led most experts to conclude that the fingerprints were not a match, thereby contradicting their previous judgments (Dror et al., 2006). Similarly, research by Elaad and colleagues (1994) looked at how prior expectations of polygraph examiners affected their decisions. One group of experts was shown a chart from a polygraph examination and told that the chart came from a suspect who had confessed. The other group of experts were shown the same chart, but were told it came from a suspect while someone else had already confessed to the crime. Results showed that the first group of experts scored the charts as more deceptive than the second group. Hill, Memon, and McGeorge (2008) examined how extra-domain information may guide hypothesis testing. In their study, participants were asked to formulate interview questions to determine whether an individual cheated on a task after being led to believe that the suspect was most likely either innocent or guilty. Participants who had been supplied with the guilty scenario asked more guilt-presumptive questions than those who had been provided with the innocent scenario. Hill et al. (2008) suggested that this was a manifestation of confirmation bias, because participants looked for information that supported their expectations. Their interviews were recorded on tape and independent observers watched the taped material. Suspects who responded to guilt-presumptive questions were judged as appearing guiltier than those who responded to questions in the innocent scenario condition.

These studies illustrate how the relevance of the issue of contextual bias within the criminal justice system is. The assumption of guilt not only influences the hypothesis testing strategies of the forensic expert, but also the assessment of statements by independent observers. Once one has categorized an individual as low in credibility, experts have a hard time considering alternative scenarios (Rassin, Eerland, & Kuijpers, 2010) and will be more sensitive to evidence that supports their expectation than to evidence that undermines it.

Following this line of reasoning, one wonders what would happen when SCAN analysts are supplied with what has been called extra-domain information about a case (Risinger et al., 2002). If SCAN is indeed sensitive to contextual bias, one would expect that such extra-domain information influences the SCAN experts' credibility judgments. In that case, the method would have a considerable error potential because not only the verbal quality of the statement would count, but also potentially unsubstantiated information.

Thus, the aim of the current experiments was twofold. First, we wanted to know whether SCAN is vulnerable to contextual or expectancy bias induced by extra-domain information. Second, we wanted to explore how SCAN fares with respect to contextual bias when it is compared to other methods of verbal credibility assessment. To this end,

Experiment 1 relied on SCAN trained police officers, while Experiment 2 evaluated in undergraduate students the liability to contextual information of two additional credibility assessment methods. In the second experiment, all participants were presented with statements that they had to analyse with one of three methods [Criteria Based Content Analysis (CBCA), Reality Monitoring (RM), SCAN or none], while they had been exposed to credibility enhancing or reducing information about the context of the statement. If these methods are sensitive to contextual bias, one would predict that a statement would be scored as more credible when preceded by credibility enhancing cues than when it is preceded by credibility reducing cues.

EXPERIMENT 1: IS SCAN SENSITIVE TO CONTEXTUAL BIAS?

5.2 METHOD

Participants

All 32 participants read and signed a letter of Informed Consent before they took part in this study. The SCAN group consisted of 16 police officers from Belgium and the Netherlands who had completed a SCAN introductory course. Four of them had also completed an advanced SCAN course. The control group consisted of 16 police officers who had never used SCAN. The mean age of the participants (9 women) was 40.6 years (*SD* = 8.3). This study was approved by the Ethics Committee of the Faculty Psychology and Neuroscience, Maastricht University

Materials

Selection of statements

To ascertain ecological validity, four statements were selected from case files of the AMS Police. Names and places in the statements were changed to protect privacy. Statements have been provided by alleged victims of different crimes (i.e., sexual abuse, rape, murder and kidnapping) and the lengths of these statements were 392, 286, 328, and 239 words, respectively. In a pilot, we tested whether the a-priori credibility of these statements was comparable. Pilot participants (*n* = 10) indicated how credible they found each statement on a 7-point scale (1 = not credible; 7= very credible). Means and standard deviations were *M* = 4.4 (*SD* = 1.71) for the sexual abuse, *M* = 3.9 (*SD* = 1. 45) for the rape, *M* = 3.9 (*SD* = 1.66) for the murder, and *M* = 3.5 (*SD* = 1.65) for the kidnapping statement. All statements were given a mean score varying between 3.5 and 4.5, which indicates no clear preference for one statement over the other in terms of credibility.

For each statement, both positive and negative context information was fabricated to enhance or reduce credibility of the statement. This context information related to

details of the crime, with positive information intended to make the statement more believable, and negative information making the statement less believable. Thus, raters' expectations about truthfulness were manipulated by supplying them with extra-domain information such as another eyewitness confirming certain details of the statements (positive information/increasing credibility), or details about the criminal background implying a history of lying (negative information/reducing credibility). This information was given before the participants read the actual statement. Appendix F provides an example of this extra-domain information.

Procedure

All participants filled in the informed consent and a short questionnaire about their work as a police officer (age, gender, and years of experience) that was used to recruit a matching control group. For the group of SCAN trained police officers (4 women), the means for age and years of experience were M = 42.13 (SD = 7.80) and M = 17.71 (SD = 11.58), respectively. For the control group (5 women), these means were M = 39.06 (SD = 8.70), and M = 15.13 (SD = 11.61), respectively. Independent samples t-tests showed no significant differences between both groups for age (t (30) = 1.05, p = 0.30), or experience (t (30) = .63, p= 0.53).

Next, the participants were given the extra-domain information and the four statements. Between participants, each statement was presented along with credibility enhancing or reducing information equally often. To exclude any order effects, the order of presentation of the statements was balanced according to a Latin square (Williams, 1949). At each of the four positions, each statement was presented once with credibility enhancing, and once with credibility reducing information, resulting in 16 unique orders, one for each participant in each group. Next, participants were asked to analyse each statement using either SCAN, or no credibility assessment method (control group). More specifically, participants in the control condition were asked to read the information and answer the subsequent question "How credible do you find this statement, based on your analysis?" on a 7-point Likert scale, ranging from 1 (not credible) to 7 (very credible). Participants in the SCAN condition were asked to first perform a SCAN analysis over the statements and then answering the following questions: "How credible do you find this statement, based on your analysis?" on a 7-point Likert scale. Following this, the SCAN group was asked to "Please write down which of the SCAN criteria you used to analyse the statement?" Because SCAN lacks a formal scoring procedure and the different criteria can be weighed differently, participants were also asked to "Please write down on which criteria you based your 7-point credibility rating?". Participants all brought their SCAN manual and were told to use the manual for criteria, when necessary. In this way, they had access to all SCAN criteria. We did not provide SCAN analysts with a list of criteria.

Inter-rater reliability

As SCAN is an unstandardized method, we first investigated which SCAN criteria the SCAN trained police officers used when analysing the statements. To this end, two independent raters coded the different criteria that were reported by the participants, using the list given in Appendix G. One rater had completed the SCAN basic course and the other rater had read the SCAN course manual and was familiar with the SCAN literature (Bockstaele, 2008a, 2008b) and colour coding scheme SCAN experts use to indicate the presence of criteria. It is important to note that the raters only coded which criteria were listed by the participants. They did not code whether they deemed the use of the colour scheme employed by the participants to be appropriate. As a result, the analyses described below cannot be interpreted as a measure of inter-rater reliability of the SCAN method. This inter-rater reliability only shows the agreement of the two raters regarding the criteria that were deemed as present by the participants.

First, the two raters scored which criteria the participants listed when answering the question "Which of the SCAN criteria did you use to analyse the statement?" Presence of a criterion was coded as '1' and absence as '0'. Criteria were coded as present if the SCAN trained police officer explicitly mentioned the criterion or articulated considerations that were in agreement with the definition of a criterion (See Appendix G). Inter-rater reliability was calculated for each criterion by dividing the number of statements where both raters agreed on the presence or absence of the criterion by the total number of statements. For example, for the *pronouns* criterion, both raters agreed on its presence or absence in the sexual abuse SCAN evaluation of 13 out of 16 evaluations. This resulted in an inter-rater reliability of 13/16=0.81. Inter-rater reliability for the coders varied for the different criteria, with a minimum of .67 and a maximum of 1. Average agreement between raters for all criteria was .90 *(SD* = 0.07*)*.

The two raters also coded participants' responses to the question on which criteria they had based their 7-point credibility ratings. Inter-rater reliability here varied between the different criteria with a minimum 0.75 and a maximum of 1. Average agreement for all criteria was .96 (*SD* = 0.06).

5.3 RESULTS

SCAN criteria

Only criteria where both raters agreed on their presence were coded as present. When raters disagreed, the criterion was coded as absent. Table 5.1 shows how many times each of the SCAN criteria were present in the statements, and how many times they were used for the credibility judgement. Six SCAN criteria were present in more than 20% of the statements. These criteria were "Pronouns" (43%), "Structure" (50%), "So-

cial introduction" (24%), "Missing time" (24%), "First person singular, past tense" (20%), and "Change in language" (27%). Criteria that were most often used to judge the statements were "Structure" (28%) and "Emotions" (21%). Furthermore, a high correlation (r = .80) was found between the criteria SCAN analysts used to analyse their statement and the criteria SCAN analysts used to make judgments about the credibility. This high correlation indicates that almost all criteria that were used to analyse the statements were also used for the credibility judgments.

Table 5.1. Number of participants who used each criterion for either their analysis or subsequent judgment in experiment 1, averaged per account.

Criteria	Used in analysis	Used for judgement
Structure of the statement	12.5	7
Use of pronouns	10.75	3.5
Change in language	6.75	1.25
Social introduction	6	0.25
Missing time	6	2.5
First person singular. past tense	5	1.5
Unimportant information	4.25	2.5
Place of Emotions	4	5.25
Unasked explanations	3.5	1.25
Objective versus subjective time	3.25	1.75
First sentence	2.25	0
Communication	1.5	0.25
Verb leaving	1.25	0.5
Exact location	0.75	0
Together with	0.75	0
Activities	0.5	0
Order	0.25	0
Out of sequence info	0	0
Extraneous information	0	0
Negative language	0	0
Resistance during rape	0	0.25

SCAN and contextual bias

To test whether the statements presented with positive context information scored higher in credibility compared with statements presented with negative context information, the credibility scores of each participant for the two statements presented along with positive context information were averaged, as were the scores of the two statements presented with negative context information. This resulted in two scores for each participant. Next, a 2 (INFORMATION: positive vs. negative) X 2 (METHOD: SCAN vs. control) mixed-model Analysis of Variance (ANOVA), with INFORMATION as a within

subject factor and METHOD as a between subject factor was conducted. Results revealed a main effect of INFORMATION, indicating that when the statements were preceded by positive context information, they were perceived as more credible (M = 4.15; SD = 1.38) than when they were preceded by negative context information (M = 2.80; SD = 0.82), (F (1, 30) = 28.25, p < 0.001; η_p^2= .49). The main effect for METHOD and the METHOD X INFORMATION interaction did not reach significance [F (1, 30) = 0.07, p = .79; d = 0.04 and F (1, 30) = 0.50, p = .49; η_p^2 = 0.016, respectively], indicating that compared with the control group, the use of SCAN did not mitigate the effects of extra-domain information on credibility ratings.

5.4 DISCUSSION

The results of experiment 1 showed that the use of SCAN did not reduce expectancy bias induced by extra-domain information, as a good forensic tool should. One could, however, argue that asking participants to indicate the credibility of a statement on a 7-point scale does not provide a high-quality measure of contextual or expectancy bias. Even though the specific instructions to the participants emphasized that they should base their judgement on their SCAN analysis ("How credible do you find this statement, based on your analysis?"), one cannot exclude that, besides basing their judgement on their SCAN analysis, participants also deliberately took into consideration the context information that is given (see for a similar line of reasoning Ben-Shakhar, Bar-Hillel, Bilu, & Shefler, 1998). In that case, participants are not exhibiting a contextual bias, but rather use information from different sources in the most optimal way. With this in mind, we carried out a second experiment to test sensitivity of SCAN and two additional verbal credibility assessment methods (CBCA and RM) to contextual bias. In the second experiment, we used a more standardized scoring system for each method, allowing us to investigate contextual bias in a more stringent way.

EXPERIMENT 2: ARE SCAN, CBCA AND RM SENSITIVE TO CONTEXTUAL BIAS?

5.5 INTRODUCTION

Experiment 1 suggested that SCAN might be susceptible to contextual bias effects. Is this also true for other, more standardised, methods of verbal credibility assessment? Apart from SCAN, at least two methods use verbal indicators for credibility assessment. The first is the Criteria Based Content Analysis (CBCA).

The CBCA was originally developed in Germany to analyse the credibility of child witness statements in sexual abuse cases. Undeutsch (1967) argued that children's statements about true events differ in content and quality from their statements about

fabricated events. Based on these differences, he developed a list of criteria to evaluate the credibility of witness testimonies. Steller and Köhnken (1989) refined these criteria and integrated them in a formal system as it is used today. CBCA is the third phase in a more extensive four-phased credibility assessment method called Statement Validity Assessment (SVA). While CBCA is a systematic analysis of the content of a particular statement, SVA is a more general credibility assessment incorporating additional information from different sources beside the statement. The first phase of this method consists of investigating all possible information about the specific case. In the second phase, the victim (witness) is interviewed about the incident. A transcript of this interview is then analysed in the third phase with the CBCA. The fourth phase includes a validity checklist for eliminating other issues that could have influences CBCA analysis (Steller, 1989; Vrij, 2008b).

Although CBCA was initially developed for evaluating children's testimonies, numerous studies have shown its usefulness with adult victim and/or eyewitnesses (Akehurst et al., 2001; Sporer, 1997; Vrij et al., 2004a; Vrij, Edward, Roberts, & Bull, 2000). A qualitative review by Vrij (2005) showed that the accuracy rate of CBCA varied between 55% and 90%, with an average accuracy rate of 70% (accuracy rates were based on observer' ratings or discriminant analyses). CBCA consists of a subset of cognitive and motivational criteria. Cognitive criteria are criteria that are likely to indicate true statements, as they are typically too difficult to fabricate (i.e., details about time and place, descriptions of interactions). On the other hand, motivational criteria refer to how the witness presents a statement. Liars are concerned about making a credible impression and therefore leave out information that may potentially damage their story (i.e., raising doubts about one's own testimony, admitting lack of memory; Vrij, 2005). When the individual cognitive and motivational criteria were considered, results of Vrij (2005) showed that the cognitive CBCA criteria had a higher diagnostic value than the motivational criteria. However, DePaulo et al. (2003) did find evidence that truth tellers included more spontaneous corrections and acknowledged their inability to remember something more than liars.

Besides CBCA, Reality Monitoring has also been shown to distinguish true from false statements. Reality Monitoring refers to the cognitive operations that a person relies upon to attribute memories to internal (fabricated) and external (perceived) events (Johnson & Raye, 1981). The rationale behind the RM method is that memories of true events will differ in quality and content from fabricated memories in a number of ways (Johnson & Raye, 1981). Since the 1990's, scientists are interested in whether RM can be used to discriminate between true and false statements (Sporer, 1997; Vrij, 2008b). A first set of proposed RM criteria were the eight criteria discussed by Sporer (1997), which reflects aspects such as realism, details about space and time, sensory information, and clarity/vividness. Studies have shown that when summing the scores of the different criteria, the average accuracy rate of RM is comparable to that of CBCA and varies between 61% and 83%, with an average of 69% (Vrij, 2008b). As to the individual

criteria, the contextual (temporal and spatial) criteria seem to have the highest diagnostic value (Masip et al., 2005).

The question to what extent CBCA, RM, and SCAN are vulnerable to contextual bias is especially relevant in light of guidelines concerning the handling of extra-domain information. As previously mentioned, unlike SCAN, SVA guidelines stress that in the first phase the expert should gather as much information as possible about the case and about the person who wrote the statement. This means that a CBCA analyst has knowledge of contextual information when analysing the statement. However, only information on the background of the victim's cognitive and verbal competence is relevant for the CBCA evaluation. The evaluation of other data (e.g., biographical information, behavioral information, etc.) is only necessary when making judgments about the complete overall credibility (Steller, 1989). In sum, a CBCA analyst has knowledge about different types of background information, other than the victim's cognitive and verbal competence. This could be considered extra-domain information, and could potentially influence the credibility assessment of the statement if CBCA is sensitive to contextual bias.

Experiment 2 tested to what extent CBCA, RM, and SCAN are sensitive to contextual bias. In addition to the 7-point scale we used in Experiment 1, in Experiment 2, we also analysed the scoring of the criteria for each method to provide a more stringent test of the sensitivity of these methods to contextual bias.

5.6 METHOD

Participants

A total of 128 undergraduate students (30 men) of Maastricht University participated in this experiment. The mean age of the participants was *M* = 22.5 years (*SD* = 5.1). Participants were randomly assigned to one of four groups; CBCA, RM, SCAN or control group. The study was approved by the Ethics Committee of the Faculty of Psychology and Neuroscience, Maastricht University.

Procedure

Participants were tested in small groups (average *n* = 4). They were seated separately from each other, to ensure that they were not able to look at each other's scores. Each group in the CBCA, RM, and SCAN conditions received a 30-minute training on how to use these assessment methods. More specifically, participants received information about the different criteria as described in chapters 8 to 10 of Vrij's book (2008). Multiple short examples were discussed to help participants to understand each criterion. After all criteria and their short examples were discussed, participants received an ex-

ample statement on which they were asked to practice the scoring of the criteria. Their codings were discussed and all questions participants still had, were answered.

In this experiment, participants were instructed to score 19 CBCA criteria, 8 RM, or 12 SCAN criteria (see Appendix G; for a detailed overview see Vrij, 2008). All participants were given the extra-domain information and the statements in the same counterbalanced order as in Experiment 1, and were asked to score each criterion indicating truthfulness on a 3-point scale (0 = absent, 1 = somewhat present, and 2 = strongly present). RM and SCAN also consist of criteria indicating deception. For RM, this was only one criterion (i.e., cognitive operations). For SCAN, there were 8 criteria that are thought to flag deception (marked with * in Appendix G). Participants were asked to reversely score these deception criteria (0 = absent, -1 = somewhat present, and -2 = strongly present). Criteria sums scores for each method were computed by summing the individual criteria. Thus, for CBCA, total scores had a possible range from 0 to 38, for RM, they had a possible range from -2 to 14, and for SCAN total scores had a possible range from -16 to 8, with a higher number indicating a higher credibility score.

After participants evaluated a statement using CBCA, RM, SCAN or no method, they rated the credibility of that statement by completing a 7-point Likert scale, ranging from 1 (not credible) to 7 (very credible). This procedure was repeated for each of the four statements.

Inter-rater reliability

To check the effectiveness of the 30- minute training, inter-rater agreement was calculated. One possible inter-rater reliability coefficient is intra-class correlation coefficients (ICC). How these coefficients are quantified is dependent on the specific design that is used to determine inter-rater reliability (see Shrout & Fleiss, 1979). Because in our design, different participants rated different combinations of statements and type of information (positive or negative context information), not all the sources of variation that must be determined to compute the ICC could be estimated from our data. We therefore did not use a simple ICC parameter to measure agreement. Instead, we used an alternative that would meet the restrictions of our design. We focused on inter-rater agreements for the 8 different 'statement x type of information' combinations. Each combination was rated by 16 participants, which permitted us to compute r_{wg}, a measure of within-group interrater agreement, developed by James, Demaree, and Wolf (1993). The r_{wg} measure has a range of 0 to 1, and indicates the proportional reduction of error variance due to agreement amongst raters. Complete agreement amongst judges would result in an observed variance equal to zero, and therefore the r_{wg} would be equal to 1. On the other hand, a total lack of agreement would result in a uniform score distribution, with an observed variance equal to the expected score variance for a uniform distribution, and a resulting r_{wg} equal to 0.

Using a uniform score distribution for computing the expected error variance of CBCA, RM, and SCAN, we found r_{wg} values for CBCA that ranged from 0.83 to 0.97, r_{wg} values for RM that ranged from 0.60 to 0.87, and r_{wg} values for SCAN that ranged from 0.67 to 0.89 (see Table 5.2.). These estimates should be interpreted with caution, as their validity hinges on the correctness of the distribution that was chosen as a model for random responding. A uniform distribution seems plausible, but any deviation from it will decrease values of r_{wg}. With this proviso in mind, we feel that our r_{wg} values suggest that the three verbal credibility assessment methods were similar in the consistency with which participants applied them to the statements after a 30-minute training, although there were differences in level of agreement, with CBCA yielding more agreement amongst observers than either RM or SCAN.

Table 5.2. Inter-rater agreements (R_{wg}) for the 8 different 'statement x type of information' combinations in experiment 2.

	Negative information				Positive information			
Method	S1	S2	S3	S4	S1	S2	S3	S4
CBCA	0.88	0.91	0.92	0.96	0.86	0.83	0.97	0.91
RM	0.87	0.73	0.60	0.65	0.81	0.72	0.73	0.68
SCAN	0.67	0.71	0.89	0.82	0.80	0.80	0.80	0.81

Note. S1, S2, S3, S4 are statement 1, 2, 3, and 4 respectively.

5.7 RESULTS

Mean credibility scores and contextual bias

As in Experiment 1, we averaged for each participant credibility ratings of the two statements presented with positive context information and credibility ratings of the two statements presented with negative context information. This resulted in two credibility scores for each participant. A 2 (INFORMATION: positive vs. negative) X 4 (METHOD: CBCA vs. RM vs. SCAN vs. control) mixed-model ANOVA on the 7-point credibility ratings revealed a main effect of INFORMATION, indicating that credibility ratings of the statements were higher when they were preceded by positive context information (M = 4.66; SD = 1.11) than when they were preceded by negative context information (M = 3.03; SD = 1.00), (F (1,124) = 150.40, $p < 0.001$; η_p^2 = 0,55). The main effect for METHOD and the METHOD X INFORMATION interaction did not reach significance [F (1, 124) = 2.19, p = 0.09; η_p^2 = 0.05 and F (1,124) = .72, p = 0.54; η_p^2 = 0.02, respectively)]. Apparently, the use of CBCA, RM or SCAN did neither increase nor decrease credibility ratings compared to the control group (see Table 5.3).

Table 5.3. Mean (M), standard deviation (SD), skewness, and standard error for the credibility scores in experiment 2 for positive and negative context information separated for each method.

	Negative		Positive	
Condition	M (*SD*)	Skewness (*SE*)	M (*SD*)	Skewness (*SE*)
Control	2.89 *(.83)*	0.37 *(.41)*	4.84 *(.94)*	-0.32 *(.41)*
CBCA	3.11 *(1.09)*	0.02 *(.41)*	4.56 *(1.14)*	-0.16 *(.41)*
RM	3.27 *(1.08)*	-0.60 *(.41)*	4.89 *(1.17)*	-.83 *(.41)*
SCAN	2.86 *(.98)*	0.06 *(.41)*	4.36 *(1.13)*	-.73 *(.41)*

Criteria scores and contextual bias

To test whether participants found statements to be richer in criteria depending on extra-domain information, the criteria sum scores for each method were analysed. The sum scores for the two statements presented with positive context information were averaged, as were the scores for the two statements presented with negative context information. Following this, we converted the scores into within-participant *Z* scores to make CBCA, RM, and SCAN scores comparable. Next, A 2 (INFORMATION: positive vs. negative) X 3 (METHOD: CBCA vs. RM vs. SCAN) mixed-model ANOVA on the Z-scores was performed. As expected, results again revealed a main effect of INFORMATION, (F (1,93) = 42.21, $p < 0.001$; $\eta_p^2 = 0.31$), showing that credibility ratings of the statements were higher when they were preceded by positive context information (M = 0.41; SD = 0.94) than when they were preceded by negative context information (M = -0.41; SD = 0.88). The main effect for METHOD and the METHOD X INFORMATION interaction did not reach significance, indicating that the criteria sum score for CBCA, RM, and SCAN did not differ in their sensitivity to contextual bias.

The unstandardized criteria sum scores for the method and information conditions are shown in Table 5.4. Additional paired samples *t*-tests showed significant differences between participants who had been supplied with positive or negative context information with regard to their CBCA scores (t (32) = 3.26, p = 0.003, d= 0.83), RM scores (t (32) = 4.54, $p < 0.001$, d= 1.18), and SCAN scores (t (32) = 3.47, p= 0.002, d= 0.73). Apparently, participants found that the statements met CBCA, RM or SCAN criteria more when they were preceded by positive than when they were preceded by negative information, which reflects a profound contextual bias effect.

Individual criteria analyses

For the interested reader, Appendix H provides an overview of the Pearson correlations between the individual CBCA, RM, and SCAN criteria, the total sum score and the associated credibility judgment. This shows to what extent the separate criteria contributed to the total sum score and to the credibility judgment. Appendix I provides a detailed overview of the influence of the contextual information on the individual CBCA, RM and

SCAN criteria. Results indicate that six CBCA criteria, six RM criteria and four SCAN criteria were significantly influenced by the contextual information.

Table 5.4. Mean (M), standard deviation (SD), Skewness and standard error (SE) of the criteria sum scores for negative and positive context information separated for the three methods in experiment 2.

	Negative		Positive	
Method	M (*SD*)	Skewness *(SE)*	M (*SD*)	Skewness (*SE)*
CBCA	10.23 *(3.28)*	.35 *(.41)*	12.98 *(3.36)*	.65 *(.41)*
RM	6.11 *(2.39)*	-1.18 *(.41)*	8.83 (*2.17)*	-.40 *(.41)*
SCAN	-3.48 *(2.98)*	-.96 *(.41)*	-1.44 (*2.59)*	.46 *(.41)*

5.8 GENERAL DISCUSSION

Are verbal credibility methods similarly sensitive to contextual bias? Based on the current experiments, we would argue that the answer is affirmative. Experiment 1 suggested that trained SCAN trained police officers exhibit a contextual bias. Their bias was no different from that in police officers who evaluated statements without SCAN. This indicates that the use of SCAN does not mitigate contextual bias, let alone that it immunizes against such bias, as a good instrument should do.

Experiment 2 investigated to what extent other assessment methods are also susceptible to contextual bias. We found that CBCA, RM, and (again) SCAN were all affected by such a bias. In all conditions, statements presented with positive context cues were judged as more credible than statements presented with negative cues. We found no difference between the control group and the groups who relied on the CBCA, RM or SCAN to evaluate statements, suggesting that these methods do little to decrease the influence of biasing context information.

As we argued in the discussion of experiment 1, the use of a 7-point credibility scale may be suboptimal for establishing sensitivity to contextual bias. For this reason, in experiment 2, we also examined to what extent CBCA, RM, and SCAN criteria were deemed present in the statements. Ideally, this should depend entirely on the statements and should be independent from other information. Statements preceded by positive context information were found to be richer in criteria than statements preceded by negative information. So, even when they analyse the very same statements, participants found more evidence for the presence of various credibility criteria when they had been exposed to positive cues, than when they had been exposed to negative cues. This, of course, comes close to how confirmation bias is defined, namely the "selective focusing on features that are compatible with a currently held hypothesis" (Shafir, 1995; p. 267). This finding is also interesting as Wegener (1989) stresses that the main purpose of credibility assessment is assessing the credibility of the statement and not the credibility of the witness, and information about the general untrustworthiness

of the witness (e.g., lying in everyday life) should not be taken into consideration for the evaluation of the specific statement. However, participants in our study used exactly these types of information to guide their credibility evaluation.

The current findings also relate to a flexible interpretation of evidence, which has been termed the "elasticity" of the evidence. As has been documented by previous studies, various categories of evidence differ in their elasticity, i.e., the extent to which they are open to subjective interpretations. Ask, Rebelius, and Granhag (2008) investigated elasticity as a potential moderator of contextual influence. Participants were given information about a homicide case, suggesting that the suspect was guilty. Next, they were presented with either consistent or inconsistent DNA, photo, or witness evidence. Participants rated the inconsistent evidence as less reliable and generated more arguments to question its reliability than the consistent evidence. This asymmetrical scepticism was stronger for participants judging witness evidence, compared to DNA and photo evidence. This shows that especially 'soft' evidence, such as witnesses are highly sensitive to contextual bias. Given that CBCA, RM, and SCAN can most likely be categorizes as 'soft', elasticity may explain their vulnerability to contextual bias.

Experiment 2 was carried out with undergraduate students who received a 30-minute training in the verbal credibility assessment method they were instructed to use. Even though the training was short, inter-rater reliability estimates suggest that the training was sufficient to apply the methods in a similar way. Contrary to the low reliability SCAN suffers from in practice (see Chapter 3), reliability in the current study was sufficient. This is due to providing participants with a fixed list of SCAN criteria, their definitions, and a fixed coding scheme (see also Chapter 4). Furthermore, as for the SCAN, Experiment 2 reproduced the contextual bias results of Experiment 1, in which the police officers had been formally trained in SCAN. From this, we may conclude that students were equally competent as experts to apply the SCAN method and that both students and experts were affected by extra-domain information in a similar way.

In sum, our experiments demonstrate that verbal credibility methods are susceptible to a contextual bias. We feel that our research highlights an important shortcoming of such instruments that is not appreciated in manuals and articles on verbal credibility methods. The straightforward lesson that can be learned from our experiments is that, when applied to statements of victims or witnesses, verbal credibility assessment method should be used without any background information that could support or dispute the statement that is assessed.

APPENDIX F

All participants received information about each of the four statements. For each statement enhancing and reducing information was fabricated. For example, for the sexual abuse case, the extra-domain information was the following:

Negative information/reduce credibility: This is a report of sexual abuse that allegedly took place several years ago. The alleged victim stated that her uncle abused her. The interrogation of the victim's mother showed that the victim has a lot of problems at school. These problems are mainly due to her rebellious and deceitful behavior toward peers and teachers. The victim also told the mother that she was raped by a friend six months ago, but later admitted that this was consensual. The relationship between mother and the victim has recently deteriorated, partly because the victim has repeatedly stolen money.

The suspect denies that the abuse has occurred. The suspect also indicated that the alleged victim probably wants to get back at him, since he has denied the girl to go into the city with her friends, and this would be her way to do so.

Positive information/increasing credibility: This is a report of sexual abuse. The alleged victim stated that she was abused by her uncle. This is not the first time he is suspected of sexual abuse. Three years ago, his former girlfriend reported that he sexually abused her 10-year-old daughter. The case was dismissed because of lack of evidence. However, the police did find child pornography on his computer.

The suspect is described as hot-tempered by several people in his neighbourhood. This description is confirmed by the mother of the alleged victim, who indicates that the suspect had always found it difficult to control his emotions.

The suspect denies having sexually abused his niece, and states that he has no idea where the accusation comes from. He reported he always had a good relationship with the alleged victim.

APPENDIX G

SCAN criteria (Vrij, 2008)

1. Denial of allegation
2. Social introduction
3. Spontaneous corrections*
4. Lack of conviction or memory*
5. Structure of the statement*
6. Emotions
7. Objective and subjective time
8. Out of sequence and extraneous information*
9. Missing information*
10. First person singular, past tense*
11. Pronouns*
12. Change in language*

APPENDIX H

Detailed overview of correlations between the individual criteria of CBCA, RM and SCAN and their total sum score (S) and credibility score (C) separated for information type.

	CBCA				RM				SCAN			
	Positive		Negative		Positive		Negative		Positive		Negative	
Criteria	S	C	S	C	S	C	S	C	S	C	S	C
1	.285	.125	.355*	.296	.745**	.689**	.381*	.501**	.159	.004	.472**	.335
2	.326	.183	.117	.472**	.574**	.358*	.374*	.089	.442*	.399*	.163	-.004
3	.342	.033	.565**	.264	.631**	.213	.739**	.516**	-.318	-.274	.396*	.219
4	.195	.324	.455**	.247	.495**	.131	.638**	.400*	.491**	.362*	.463**	.249
5	.485**	.380*	.396*	.232	.447*	.307	.396*	.067	.700**	.572**	.388*	.512**
6	.565**	.190	.504**	.398*	.509**	.148	.750**	.446*	.679**	.487**	.590**	.309
7	.306	.111	.372*	.349*	.674**	.589**	.732**	.583**	.525**	.187	.575**	.389*
8	.230	.143	.477**	.161	.308	.455**	.474**	.477**	.540**	.423*	.523**	.473**
9	.577**	.164	.439*	.206					.496**	.523**	.519**	.307
10	.267	.278	.235	-.206					.224	.130	.331	.133
11	.487**	.260	.420*	.150					.643**	.313	.637**	.604**
12	.625**	.527**	.608**	.389*					.319	.278	.383*	.185
13	.288	.133	.340	.127								.
14	.494**	.228	.612**	.190								
15	-.079	-.225	.533**	.332								
16	.348	.248	.309	.094								
17	.309	.151	.143	-.083								
18	.296	.115	.098	-.353*								
19	.371*	.059	.414*	.048								

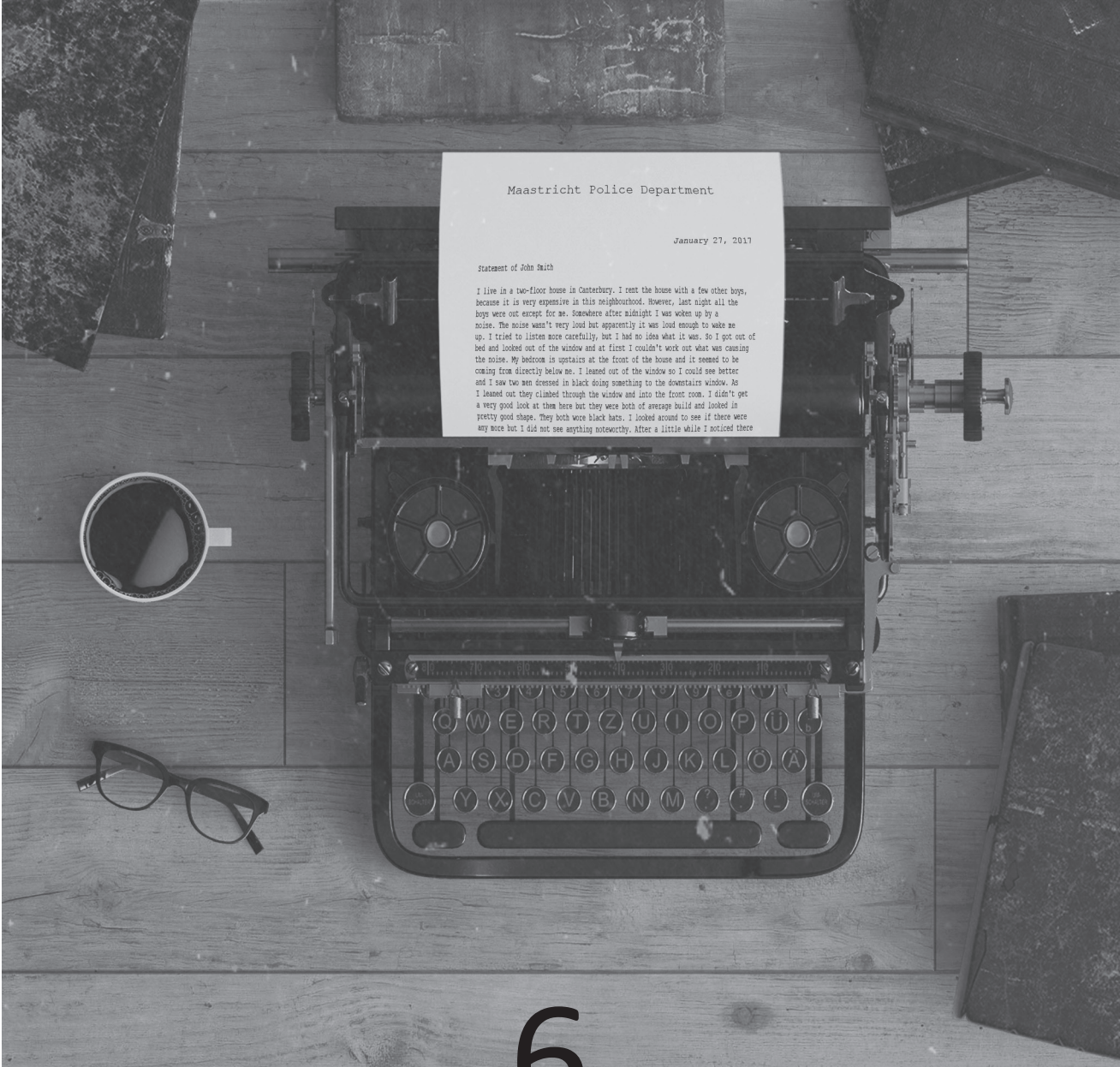
Note. **. Correlation is significant at the 0.01 level (2-tailed). *. Correlation is significant at the 0.05 level (2-tailed). Numbers of criteria refer to the numbers in appendix C.

APPENDIX I

Means (M) and standard deviations (SD) of the score for each criterion of CBCA, RM and SCAN as a function of information type (Positive vs. Negative)

Method	Criteria	Positive		Negative			Effect
		M	SD	M	SD	*t*	size (r)
CBCA	1	1.53	0.44	1.33	0.49	2.08*	0.21
	2	0.64	0.61	0.28	0.58	2.62*	0.28
	3	1.53	0.38	1.23	0.44	3.32*	0.34
	4	1.25	0.52	0.98	0.57	1.92	0.24
	5	1.39	0.59	1.3	0.47	0.77	0.08
	6	1.02	0.63	0.86	0.56	0.87	0.13
	7	0.41	0.43	0.22	0.38	2.04*	0.23
	8	0.77	0.54	0.53	0.44	2.18*	0.24
	9	0.73	0.61	0.61	0.52	0.96	0.11
	10	0.13	0.25	0.06	0.21	1.28	0.15
	11	0.55	0.59	0.27	0.31	2.68*	0.28
	12	0.91	0.64	0.78	0.55	0.64	0.11
	13	0.47	0.49	0.36	0.41	0.98	0.12
	14	0.48	0.5	0.39	0.49	0.86	0.09
	15	0.45	0.43	0.36	0.44	0.85	0.1
	16	0.09	0.24	0.06	0.17	0.57	0.07
	17	0.06	0.21	0.05	0.15	0.33	0.03
	18	0.06	0.28	0.02	0.09	0.9	0.1
	19	0.52	0.59	0.55	0.51	-0.27	-0.04
RM	1	1.42	0.46	1.03	0.47	3.65*	0.39
	2	1.28	0.54	0.88	0.44	4.61*	0.38
	3	1.72	0.36	1.44	0.59	2.88*	0.28
	4	1.3	0.54	1.13	0.61	1.1	0.15
	5	0.95	0.61	0.72	0.62	1.17	0.18
	6	1.47	0.49	1.11	0.49	3.13*	0.34
	7	1.2	0.54	0.75	0.44	4.01*	0.41
	8	-0.52	0.47	-0.94	0.61	2.93*	0.35
SCAN	1	0.3	0.62	0.25	0.55	0.62	0.04
	2	0.89	0.52	0.77	0.44	1.16	0.12
	3	-0.2	0.36	-0.34	0.48	1.22	0.16
	4	-0.34	0.43	-0.59	0.57	2.37*	-0.49
	5	-1.16	0.59	-1.45	0.48	2.12*	0.26
	6	0.7	0.55	0.41	0.51	2.21*	0.26
	7	0.73	0.58	0.48	0.55	1.83	0.22
	8	-0.69	0.55	-0.98	0.62	2.51*	0.28
	9	-0.92	0.67	-1.03	0.68	0.98	0.08
	10	-0.11	0.28	-0.23	0.42	1.35	-0.68
	11	-0.42	0.44	-0.58	0.54	1.62	0.16
	12	-0.22	0.31	-0.38	0.46	1.97	0.19

Note. * indicates that p<0.05 (2-tailed). Numbers of criteria refer to the numbers in appendix G). t refers to the t-value of the difference between scores in the Positive and Negative information condition.



6

Using an example statement increases information but does not increase accuracy of CBCA, RM, and SCAN

ABSTRACT

Verbal credibility assessment methods are frequently used in the criminal justice system to investigate the truthfulness of statements. Three of these methods are Criteria Based Content Analysis (CBCA), Reality Monitoring (RM), and Scientific Content Analysis (SCAN). The aim of this study is twofold. First, we investigated the diagnostic accuracy of CBCA, RM, and especially SCAN. Secondly, we tested whether giving the interviewee an example of a detailed statement can enhance the diagnostic accuracy of these verbal credibility methods. To test the latter, two groups of participants were requested to write down one true and one fabricated statement about a negative event. Prior to this request, one group received a detailed example statement, whereas the other group received no additional information. Results showed that CBCA and RM scores differed between true and fabricated statements, whereas SCAN scores did not. Giving a detailed example statement did not lead to better discrimination between truth tellers and liars for any of the methods, but did lead to the participants producing significantly longer statements. The implications of these findings are discussed.

Published as:
Bogaard, G., Meijer, E., & Vrij, A. (2013). Using an example statement increases information but does not increase accuracy of CBCA, RM, and SCAN. *Journal of Investigative Psychology and Offender Profiling*. *11*, 151-163. doi: 10.1002/jip.1409

6.1 INTRODUCTION

Research has shown that in general, people are not good at detecting lies, and that even trained police officers generally perform only just above chance level (Aamodt & Custer, 2006; C. F. Bond & DePaulo, 2006, 2008; Vrij, 2008b). To overcome this problem, several credibility assessment methods have been developed to assist in lie detection. One group of methods is based upon the assumption that deception is reflected in language. Criteria Based Content Analysis (CBCA; Steller & Köhnken, 1989), Reality Monitoring (RM; Johnson & Raye, 1981) and Scientific Content Analysis (SCAN; Sapir, 2005) are well-known examples.

Ample evidence exists showing that CBCA and RM perform above chance level at discriminating between truthful and fabricated accounts. Yet for SCAN only a few studies exist examining its diagnostic accuracy, and these studies typically fail to find an effect (see below, Nahari et al., 2012; Porter & Yuille, 1996). In this study, we examined to what extent CBCA, RM, and especially SCAN can discriminate between truthful and fabricated accounts. In addition, we examined whether providing the participants with a detailed example statement would increase the diagnostic accuracy of these methods.

CBCA is based on the Undeutsch hypothesis, which states that statements about true events differ in content and quality from statements that are the result of imagination (Undeutsch, 1967). The German psychologist Udo Undeutsch was the first to present his so-called reality criteria to aid in the credibility assessment of statements. Steller and Köhnken (1989) took these criteria one step further and generated a set of 19 different CBCA criteria. Even though CBCA was originally developed to assess the credibility of children's statements in cases of alleged sexual abuse (Undeutsch, 1967), research has demonstrated that CBCA can also be used with adults (Blandon-Gitlin, Pezdek, Lindsay, & Hagen, 2009; Porter & Yuille, 1996; Steller & Köhnken, 1989). Research has shown that truthful statements are richer in these criteria than fabricated statements. Vrij (2005) investigated 37 studies about CBCA and concluded that CBCA is able to correctly classify 55% to 90% of the statements. The average classification accuracy in the investigated studies was 74.63%. In 2008, Vrij again investigated the accuracy for CBCA in 27 studies and found an average accuracy rate of 70.47%.

Reality Monitoring (RM) has also been employed to discriminate truthful from fabricated statements. Reality Monitoring originally refers to the mechanism that is used to distinguish between two types of memory, external and internal memories (Johnson & Raye, 1981). External memories are based on true experiences and are expected to contain sensory information such as details about smell, taste, sound, and sight. In addition, true memories contain descriptions of context information and affective details. As a whole, they are expected to be sharp and vivid. In contrast, internal memories are based on imagined events and consequently consist of cognitive operations. These operations describe inferences (such as reasoning, search, decision, and imagery processes) made by the participants during the event or when describing the event later.

For example, "It appeared to me that she didn't know the layout of the building" (Johnson, Raye, Foley, & Foley, 1981; Vrij, 2008b). As a result, these imagined memories are less detailed compared to true memories. Masip, Sporer, Garrido, and Herrero (2005) reviewed the scientific RM deception literature and found that visual and auditory details, contextual information, time information, and realism were the most discriminative criteria. Results about the cognitive operations were mixed. The RM classification was 68.80%, which is similar to that of CBCA (Vrij, 2008b).

Scientific Content Analysis (SCAN) was developed by former Israeli polygraph examiner Avinoam Sapir (2005). Typically, a SCAN analysis starts with asking the suspect, witness, or alleged victim to write down 'everything that happened' during a particular time frame. This account is referred to as a 'pure version' of the event, as it has to be produced without the presence and interference of an investigator to minimize investigator influences. There is no standardized set of SCAN criteria, but a set of 12 criteria has been used in research (Bogaard et al., 2014b; see Chapter 3; Nahari et al., 2012). Examples of criteria include whether pronouns are avoided or whether all individuals mentioned in the statement are introduced properly. The outcome of this analysis can be used to make a credibility judgement about the statement (Sapir, 2005). To date, SCAN was examined in only two experimental studies and neither experiment found evidence for SCAN. Porter and Yuille (1996) examined three SCAN criteria, structure of the statement; missing information, and first person singular, past tense. Results indicated no differences between true and fabricated statements regarding these three criteria. In a more recent study, Nahari, Vrij, and Fisher (2012) also found SCAN scores not to differ between truth tellers and liars. Interestingly, the same statements were also analysed with RM and liars and truth tellers could be discriminated based on the RM results. Despite the lack of empirical evidence, SCAN is increasingly used for lie detection purposes in countries all over the world, such as Australia, Canada, Mexico, Israel, the Netherlands, Belgium, South Africa, UK, US, Qatar, and Singapore (Retrieved from http://www.lsiscan.com/id29.htm).

Research has shown that the accuracy of verbal credibility methods is sensitive to certain manipulations. For example, research investigating 'coaching' indicates that providing participants with specific information about the rationale and the different criteria affects the CBCA and RM results (Caso et al., 2006; Vrij et al., 2002, 2004b; Vrij et al., 2000). Vrij, Kneller, and Mann (2000) found that liars who were informed about CBCA criteria obtained higher CBCA scores than uninformed liars and that the CBCA scores of informed liars did not differ significantly from truth tellers' CBCA scores. Similar results were found regarding coached liars and RM scores (Vrij et al., 2004b). The effect of coaching on SCAN evaluations has not been examined yet, but there is no theoretical reason why SCAN would not be vulnerable to coaching in the same way and to the same extent as CBCA and RM.

Besides coaching, researchers have also examined the influence of the type of interview on verbal credibility methods. For example, Köhnken, Schimossek, Aschermann,

and Höfer (1995) examined the effect of using the cognitive interview (CI) as an alternative to the structured interview on CBCA. Several studies have shown that the CI indeed led participants to give more correct details but also, to a lesser yet significant extent, more incorrect details (Fisher, Amador, & Geiselman, 1989; Geiselman, Fisher, MacKinnon, & Holland, 1986; Köhnken, Milne, Memon, & Bull, 1999; Memon, Meissner, & Fraser, 2010). Based on this memory enhancing effect, Köhnken et al. (1995) hypothesized that CI would influence the ability of CBCA to discriminate between true and fabricated accounts. Participants were asked to give a truthful or fabricated account of a blood-donation while being interviewed with either the structured interview or the CI. Statements of participants who were interviewed with the CI contained significantly more details misunderstood, lack of memory, self-doubts, unusual details and superfluous details than the statements of participants interviewed with the structured interview. Yet, CBCA scores did not differ between truthful and fabricated accounts depending on the type of interview, showing that using the CI as an alternative to the structured interview did not influence the accuracy of the CBCA.

In the light of the methods' sensitivities to manipulations, we tested whether giving individuals an example of a detailed statement, fulfilling all relevant criteria of the CBCA, RM and SCAN, without explaining all the separate criteria would influence the methods' scores. Like Köhnken et al. (1995), the purpose of this example was to obtain a more comprehensive statement from participants. However, instead of asking memory-enhancing questions, we merely gave participants an example of the result we were interested in, i.e., a more detailed statement. Supplying individuals with a more detailed statement can yield different effects. It could result in liars obtaining higher scores on the different methods to the extent that their scores would not differ anymore from those of truth tellers, like the coaching effect. However, in coaching experiments liars were informed how the veracity methods worked and which criteria they should include. Such specific instructions were not given in the present experiment. Without the exact knowledge of the different criteria within the example statement, it could become difficult for liars to produce a statement that is comparable to the example statement and of the same quality of a truth teller's statement. For truth tellers, the example statement could encourage them to include more information. The detailed example could act as a memory enhancing strategy, resulting in a more detailed statement. We therefore hypothesized that participants who received the example would give more comprehensive statements, resulting in higher scores on CBCA, RM and SCAN. In addition, we hypothesized that this effect would be more pronounced in truth tellers than in liars, thereby increasing discriminability of these methods. As mentioned earlier, besides investigating the effect of the example statement, we also investigated to what extent these methods – and especially SCAN - can discriminate between true and false statements.

6.2 METHOD

Participants

A total of 64 participants (28 males), aged 17 to 27 years old (*M* = 21.09, *SD* = 2.27) took part in the study. Mother tongue Dutch was a prerequisite. This study was approved by the Ethics Committee of the faculty of Psychology and Neuroscience of Maastricht University.

Materials

Verbal credibility methods

Criteria included in the example statement

Although CBCA originally consists of 19 criteria, only nine criteria were included in the example statement. We left out the motivational-related criteria, and the specific criteria "unexpected complications during the incident" and "related external associations" because of weak empirical support in previous studies (Vrij, 2005). Furthermore, the criterion 'accurately reported details misunderstood' was not suitable for the current study as it refers to typical child witness issues. Finally, three criteria were omitted because they had contradicting interpretations depending on the method used. These particular criteria were 'Spontaneous corrections' (both CBCA and SCAN), 'Admitting lack of memory' (CBCA) *vs.* 'Lack of conviction or memory' (SCAN) and 'Unstructured production' (CBCA) *vs.* 'Out of sequence and extraneous information' (SCAN). For CBCA, the presence of these criteria indicates truthfulness while for SCAN the presence of these criteria indicates deception. As a result, the following nine CBCA criteria were of use to this study (for a complete description of all criteria see Appendix J): (1) Logical structure, (2) Quantity of details, (3) Contextual embedding, (4) Descriptions of Interactions, (5) Reproduction of conversation, (6) Unusual details, (7) Superfluous details, (8) Accounts of subjective mental state and (9) Attribution of perpetrator's mental state (Steller & Köhnken, 1989; Vrij, 2008b). All criteria are expected to be more present in truthful statements than in deceptive statements.

We wanted to create an example statement that would be perceived as highly credible. We therefore only included in the statement criteria that indicate truthfulness and left out criteria that indicate deception. For the RM method, this means we excluded the criterion "cognitive operations". As a result, seven criteria were included in the example statement: (1) Clarity, (2) Perceptual information, (3) Spatial information, (4) Temporal information, (5) Affect, (6) Reconstructability of the story, and (7) Realism. All seven criteria are expected to be more present in truthful statements. (Sporer, 1997; Vrij, 2008b).

For the SCAN method, 6 instead of the 12 criteria reported by Vrij (2008b) were used. Three overlapping criteria mentioned above were excluded, as well as the criteri-

on "Denial of allegations" as – given the nature of the statements - it is not appropriate for the current study (participants were not accused of anything and therefore had no reason to deny the allegations). Furthermore, as mentioned before, we also excluded the criteria indicating deception, as they could not be incorporated in the example statement. The resulting six criteria included were (1) Social introduction, (2) Structure of the statement, (3) Emotions, (4) Objective and subjective time, (5) First person singular, past tense, and (6) Pronouns. See appendix J for the complete lists and definitions of the criteria.

Example statement

The example statement was produced by one additional participant who read all the necessary literature about CBCA, RM, and SCAN and received a 1.5-hour training in the three methods. To make sure the statement was detailed enough, the participant was first interviewed about the event, a recent surprise helicopter flight in New York. Next, the participant was asked to write down a very detailed (true) statement about the event, and to avoid including criteria that indicate deception. The statement consisted of 1291 words, which is approximately three pages in writing, and aimed to fulfil all criteria of CBCA, RM, and SCAN indicating truth, as described above.

To check whether each separate criterion indicating truthfulness was indeed present in the example statement, a pilot study with 12 participants (7 males, M = 20.75 years, SD = 2.45 years) was conducted. Four participants analysed the example statement with CBCA, four with RM and four with SCAN. All participants received a 1 hour training in the method they were going to use, and scored all criteria indicating truthfulness on a 3-point scale including 0 (not present), 1 (present), and 2 (strongly present; Blandon-Gitlin et al., 2009; Gödert et al., 2005). As we instructed the writer of the story not to include criteria indicating deception, we did not further investigate these criteria.

Table 6.1 shows an overview of the different scores for each criterion separated for each method. For each method, we calculated the total method score averaged for the four raters. For CBCA, the four raters obtained an average total score of 11.75. As can be seen in Table 6.1, criterion (9) "Attribution of perpetrator's mental state" was scored as absent in the example statement. This is due to the positive nature of the statement, where no perpetrator was present. For RM, the four raters obtained an average score of 10. None of criteria were scored as absent. Finally, for SCAN, the four raters obtained an average score of 8. The criterion (4) "objective and subjective time" was scored as present (1) by two raters, and as absent by the other two. We believe that these scores indicate that the different criteria were sufficiently presented in the example statement.

Procedure

Participants were told that the study was about the discriminability of verbal lie detection methods. Participants were asked to think for a maximum of three minutes about a

real and a fabricated story. They were told that both stories should be about a recent negative event in their life, either emotionally or physically. It was stressed that the fabricated story should consist of a realistic scenario that never actually occurred in their life. Furthermore, half of the participants received the 'example statement', and were told that they had to produce a statement in similar style to the example statement about the event they experienced and about an event they completely fabricated. The other half of the participants were merely asked to produce the two statements as detailed as possible. All participants were told that their statements would be used to test deception detection methods and that they had to try to write their stories as convincing as possible. The order in which the participants wrote the two statements was counterbalanced. Participants were reassured that their stories would be treated confidentially and anonymously. No time limit was given for de production of the statements.

Table 6.1. Detailed overview of the scores from the different raters for the pilot statement

		Criteria								
Method	Participant	c1	c2	c3	c4	c5	c6	c7	c8	c9
CBCA	1	2	2	2	1	2	2	0	1	0
	2	2	2	1	1	2	1	2	2	0
	3	2	2	2	1	1	0	1	2	0
	4	1	2	2	1	2	0	1	2	0
RM	5	2	2	2	2	1	1	2	-	-
	6	2	1	1	1	2	1	2	-	-
	7	2	1	2	2	2	1	2	-	-
	8	2	2	2	2	2	1	2	-	-
SCAN	9	1	2	2	0	2	2	-	-	-
	10	1	2	2	1	2	2	-	-	-
	11	1	2	2	0	2	2	-	-	-
	12	0	2	2	1	2	2	-	-	-

Note. The numbers of the criteria refer to the numbers of the criteria mentioned in the Appendix J. For each method, the numbers refer to different criteria.

Statement coding

After participants finished writing their statements, the statements were analysed with SCAN, RM, and CBCA by two independent raters for each method. All six raters were females with an average age of 22.8 years. The raters were made familiar with the literature discussing one of the three different verbal analysing methods they were expected to use (Masip et al., 2005; Nahari et al., 2012; Steller & Köhnken, 1989). Participants also received the appropriate pages of the different chapters of Vrij (2008) about CBCA (Chapter 8; 207-213), or RM (Chapter 9; 266-269) or SCAN (Chapter 10; 282-287).

In addition, the raters received a 1.5-hour training about the method they were expected to use. During the training, all the criteria were discussed separately and examples were given. After all criteria were discussed, raters received two practice statements and these were also discussed. After this training was completed, the coders started analysing the 128 written statements of this study. Two raters analysed the statements with CBCA, two raters with RM and two raters with SCAN. Raters coded the presence of all criteria, including those that were excluded from the example statement ('Spontaneous corrections', 'Admitting lack of memory', and 'Unstructured production' for CBCA, 'Cognitive operations' for RM, and 'Lack of conviction or memory', 'Out of sequence and extraneous information', 'Out of sequence and extraneous information', 'Missing information', and 'Change of language' for SCAN). Thus, for CBCA the raters used a list of 12 criteria, for RM eight criteria, and for SCAN eleven criteria.

Again, all criteria were scored on a 3-point scale ranging from 0 (not present) to 2 (strongly present). However, criteria indicating deception, indicated with an asterix in the appendix, were scored in reverse, ranging from -2 (strongly present) to 0 (not present).

Inter-rater reliability

For each separate criterion, inter-rater reliability was calculated by means of Cohen's Kappa. For CBCA, values varied between 0.47 to 0.91 with an average Cohen's Kappa of 0.70. For RM, values varied between 0.32 and 0.90 with an average Cohen's Kappa of 0.53. Finally, for SCAN, values varied between 0.49 and 0.94 with an average Cohen's Kappa of 0.72. These results indicated that there is moderate (RM) to high agreement (CBCA, SCAN) between the two raters. Although the scores are sometimes low for individual criteria, we do not think this is to damaging as we only used total scores in our analyses. Previous studies in lie detection have used similar reliabilities, and obtained even lower reliabilities than the ones found in the current study (Blandon-Gitlin et al., 2009; Vrij, 2005).

6.3 RESULTS

Statement type

Statements were coded as either an emotional or injury statement. A statement was coded as physical when the writer wrote about a physical injury that s/he experienced her/himself. All other statements were coded as emotional. For example, a statement describing the grandfather being treated for cancer would be coded as emotional. A statement describing a car crash in which the writer was hurt would be coded as an injury description. For true statements, 53 statements discussed an emotional event, and 11 statements pertained to an injury event. For the fabricated statements, 50

statements consisted of an emotional event, and 14 statements discussed an injury event. The ratio of type of statement did not differ significantly between truthful and fabricated statements [χ^2 (1, *N*=64) = n.c., *p*=0.66]. However, overall, participants seemed to prefer to write about an emotional event.

Table 6.2. Detailed overview of the Means and Standard Deviations for each SCAN criterion as a function of veracity.

	True		Fabricated		
Criteria	M	SD	M	SD	Cohen's d
Social Introduction	0.70	0.71	0.83	0.63	-0.19
Spontaneous corrections	-0.42	0.58	-0.38	0.55	-0.07
Lack of conviction or memory	-0.03	0.18	0	0	-
Structure of the statement	0.78	0.61	0.63	0.62	0.24
Emotions	0.70	0.58	0.69	0.64	0.02
Objective and Subjective time	0.30	0.42	0.31	0.42	-0.02
Out of sequence and extraneous info	-0.25	0.38	-0.25	0.38	0
Missing information	-0.25	0.40	-0.27	0.44	0.05
First person sing, past tense	1.70	0.54	1.55	0.53	0.28
Pronouns	1.67	0.45	1.72	0.46	-0.09
Change in language	-0.09	0.24	-0.2	0.42	0.33

Verbal Credibility Methods

To exclude any potential effect of the example statement, the analyses regarding the diagnostic accuracy of the three methods were only carried out on the control group. For CBCA, the raters coded 12 criteria. The other seven criteria of CBCA were excluded because of weak empirical support or because they were not suitable for the current study. For RM, the raters coded all 8 criteria. For SCAN, due to the nature of the statements, the criterion 'Denial of allegation' was not coded, resulting in 11 included criteria.

We calculated the total score for each method, by summing up 12 criteria for CBCA, 8 criteria for RM, and 11 for SCAN. True statements included significantly more CBCA criteria (*M* = 7.19, *SD* = 2.09) than fabricated statements (*M* = 6.17, *SD* = 2.16) [*t*(31) = 2.21, *p* = 0.04]. True statements also included significantly more RM criteria (*M* = 5.70, *SD* = 2.03) than fabricated statements (*M* = 4.53, *SD* = 2.10) [*t*(31) = 2.53, *p* = 0.02]. However, SCAN scores did not differentiate between true (*M* = 4.81, *SD* = 1.63) and fabricated statements (*M* = 4.62, *SD* = 1.94) [*t*(31) = 0.50, *p* = 0.62]. To allow comparison with other studies, we calculated Cohen's *d* for the main effect of veracity for the different methods. The *d*-values were 0.48 for CBCA, 0.56 for RM, and 0.11 for SCAN.

A non-significant effect for the SCAN total score does not rule out that some individual SCAN criteria discriminated significantly between truth tellers and liars. However, as Table 6.2 shows, not a single SCAN criterion significantly discriminated between true and fabricated statements.

Effect of the example statement

To investigate to what extent the example statement influenced the number of criteria present in the statements, we only included criteria that were present in the example statement in the analysis. Total CBCA score was calculated by summing up the 9 CBCA criteria, total RM score by summing up the 7 RM criteria, and total SCAN score by summing up the 6 SCAN criteria that were present in the example statement (see methods section). Mean scores and standard deviations of the three methods are presented in Table 6.3. A 2 (Veracity: truth vs. fabricated) x 2 (Example: yes vs. no) mixed ANOVA on the total CBCA scores revealed a significant main effect for Veracity [F (1, 62) = 12.52, p < .001, η_p^2 = 0.17] with true statements receiving a higher CBCA score than fabricated statements. Results also showed a significant main effect for Example [F (1, 62) = 19.65, p < .001, η_p^2 = 0.24] with participants who received an example writing a statement that scored higher on CBCA. There was no significant Veracity X Example interaction [F (1, 62) = .23, p = .63, η_p^2 = 0.01], indicating that the example statement did not significantly influence the discriminability of the CBCA.

For RM, the results of the mixed ANOVA yielded a significant main effect for Veracity [F (1, 62) = 11.71, p <.001, η_p^2 = 0.16], with true statements scoring higher on RM than fabricated statements. As well, a significant main effect for Example [F (1, 62) = 24.52, p < .001, η_p^2 = 0.28] emerged, with participants who received an example writing a statement that scored higher on RM. Again, there was no significant interaction between Veracity and Example [F (1, 62) = .15, p = .70, η_p^2 = 0.002], indicating that the example statement did not significantly influence the discriminability of the RM.

For SCAN, the mixed ANOVA showed no significant main effect for Veracity [F (1, 62) = 0.35, p = .556, η_p^2 = 0.006], revealing that SCAN did not accurately discriminate between true and fabricated statements. However, there was a significant main effect for Example, [F (1, 62) = 7.15, p = .0.01, η_p^2 = 0.10], indicating that the example statement caused participants to attain a higher SCAN score. Finally, no significant interaction effect between Veracity and Example [F (1, 62) = .005, p = 0.94, η_p^2 = 0] implying that the example statement did not significantly influence the discriminability of the SCAN scores.

Table 6.3. Means and Standard Deviations for total scores on CBCA, RM and SCAN as a function of condition for the criteria included in the example statement

	CBCA		RM		SCAN	
Condition	True	False	True	False	True	False
Explanation	8.50 (*2.71*)	7.27 (*1.81*)	7.77 (*2.42*)	6.88 (*2.04*)	6.67 (*1.55*)	6.56 (*1.58*)
Control	6.48 (*1.77*)	5.55 (*1.90*)	5.72 (*2.04*)	4.56 (*2.08*)	5.85 (*1.23*)	5.72 (*1.61*)
Total	7.49 (*2.49*)	6.41 (*2.04*)	6.76 (*2.45*)	5.72 (*2.35*)	6.27 (*1.44*)	6.14 (*1.63*)

Number of words

Means and standard deviations of the number of words present in the statements are presented in Table 6.4. Participants in the example condition wrote on average 296.78 (*SD* = 98.76) words, while participants in the control condition wrote on average 181.94 (*SD* =55.75) words. A 2 (Veracity) x 2 (Example) mixed design analysis yielded no main effect for Veracity, [F (1, 62) = 2.01, p = .16, η_p^2 = 0.03]. However, results did show a significant main effect for Example, F (1, 62) = 37.38, p < .001, η_p^2 = 0.38, indicating that participants in the example statement condition wrote a longer statement compared to the control group. No significant interaction between Veracity and Example emerged [F (1, 62) = .003, p = .96, η_p^2 = < 0.01], meaning that receiving an example statement influenced true and false in a similar way (see Table 6.3).

Table 6.4. Means and Standard deviations for the number of words in true and fabricated statements for both conditions

	True		Fabricated	
Condition	M	SD	M	SD
Explanation	301.78	102.66	291.78	96.08
Control	187.34	58.79	176.53	52.91
Total	244.56	101.06	234.16	96.40

6.4 DISCUSSION

The aim of the present study was to test the ability of CBCA, RM, and SCAN to differentiate between true and fabricated accounts. Furthermore, we investigated the effect of supplying participants with a detailed example statement on verbal credibility assessment scores. Both CBCA and RM scores were significantly higher for true compared with false statements. In contrast, SCAN failed to discriminate between true and false statements. Further investigation of the separate SCAN criteria also failed to find supporting evidence for its ability to discriminate between truthful and lied statements.

These findings are in line with the previous studies of Nahari et al. (2011) and Porter & Yuille (1996) that failed to find support for SCAN. Nahari et al. (2011) investigated SCAN total score and compared these scores to RM total scores. As in our study, RM was able to accurately discriminate between true and fabricated statements, while SCAN performed poorly in discriminating between these statements. Although the lack of significant findings for SCAN is in line with previous research, one limitation deserves attention. First, as part of their training, the SCAN raters received the article of Nahari et al. (2011), which concluded that SCAN was not able to accurately discriminate between true and false statements. This information may have influenced the motivation of the raters and thereby the quality of the analysis of the statements.

Supplying participants with an example statement resulted in higher CBCA, RM, and SCAN scores for both fabricated and truthful accounts. This lack of increased diagnostic accuracy with the example statement for CBCA is in line with Köhnken et al. (1995), who showed that using CI resulted in an increased use of several CBCA criteria. Participants who received the example statement did produce longer statements than participants who did not receive the example. As such, although an example statement did not improve the ability to detect deceit, providing interviewees with an example may still be useful as statements produced after receiving an example are likely to contain more information. As obtaining information from suspects is often seen as the core of interviewing (Bull, 2010; Fisher, 2010), proving an example statement seems to help in accomplishing this aim. The obtained information can in turn be checked and validated and may provide more insight into the suspect's involvement in the crime.

To maintain ecological validity, we deliberately did not explain the criteria to the participants (as is done in coaching studies), but hypothesized that truth tellers would implicitly incorporate more criteria from the example statement in their statement than liars. Importantly, besides the main effect on the number of words, there was also a main effect for all three method scores, showing that participants who received the example statement did include more criteria in their statement. Thus, even though the participants may not have been explicitly aware of the criteria, this shows that our example statement manipulation was successful and that the participants did not simply write a longer statement.

In the current study, participants were free to choose the events on which they based their statements. This resulted in a variety of events described, and consequently, the veracity of the statements (i.e., ground truth) could not be verified. It cannot be ruled out that some of the participants failed to follow instructions, which in turn could have influenced our results. Although it could be argued that this is a limitation of the design of this study, it is also important to realize that such a variety of events described is in fact what investigators are faced with in practice. Thus, it can also be argued that by scarifying some degree of control such a design increases external validity.

The findings of the current study have two practical implications. First, as our findings replicate that SCAN criteria do not discriminate between true and fabricated statements, the extensive use of SCAN in practice should be discouraged. Second, our data show that providing participants with an example statement results in higher CBCA, RM, and SCAN scores for both truthful and fabricated statements. This means that when using such a manipulation in practice, one should be cautious about veracity verdicts based on a high score, as these high scores may simply be due to the example statement.

In conclusion, results clearly indicate that the lie detection methods differ in their sensitivity to measure truthfulness as RM and CBCA, but not SCAN, could accurately discriminate between true and fabricated statements. Supplying individuals with a detailed example statement did not influence the discriminability of the verbal credibility assessment, but did increase the length of the statement, which might be useful for investigative purposes.

APPENDIX J (ADAPTED FROM VRIJ (2008))

Criteria Based Content Analysis (CBCA)

1. *Logical structure*: The statement should be coherent and does not contain logical inconsistencies or contradictions
2. *Quantity of details*: The statement should be rich in detail and include specific descriptions of place, time persons, objects and events.
3. *Contextual embedding*: The events should be placed in time and locations, and the actions are connected with other daily activities and/or customs.
4. *Descriptions of Interactions*: The statement should contain information that interlinks at least the alleged perpetrator and witness.
5. *Reproduction of conversation*: The statement should report parts of the conversation in the original form or the different speakers are recognizable in the reproduced dialogues.
6. *Unusual details*: Refers to details of people, objects, or events that are unique, unexpected or surprising but meaningful in the context.
7. *Superfluous details*: Refers to details that are in connection with the allegations that are not essential for the accusation.
8. *Accounts of subjective mental state*: Refers to the witness describing the development and changes of his or her feelings at the time of the incident.
9. *Attribution of perpetrator's mental state*: Refers to the witness describing the perpetrator's feelings, thoughts, or motives during the incident.
10. *Unstructured production*: The information in the statement is presented in a non-chronological order.
11. *Spontaneous corrections*: The statement contains corrections or information is added to previously provided information, without any interference of the interviewer.
12. *Admitting lack of memory*: the statement contains information admitting lack of memory such as "I don't know" or "I don't remember what happened after he hit me".

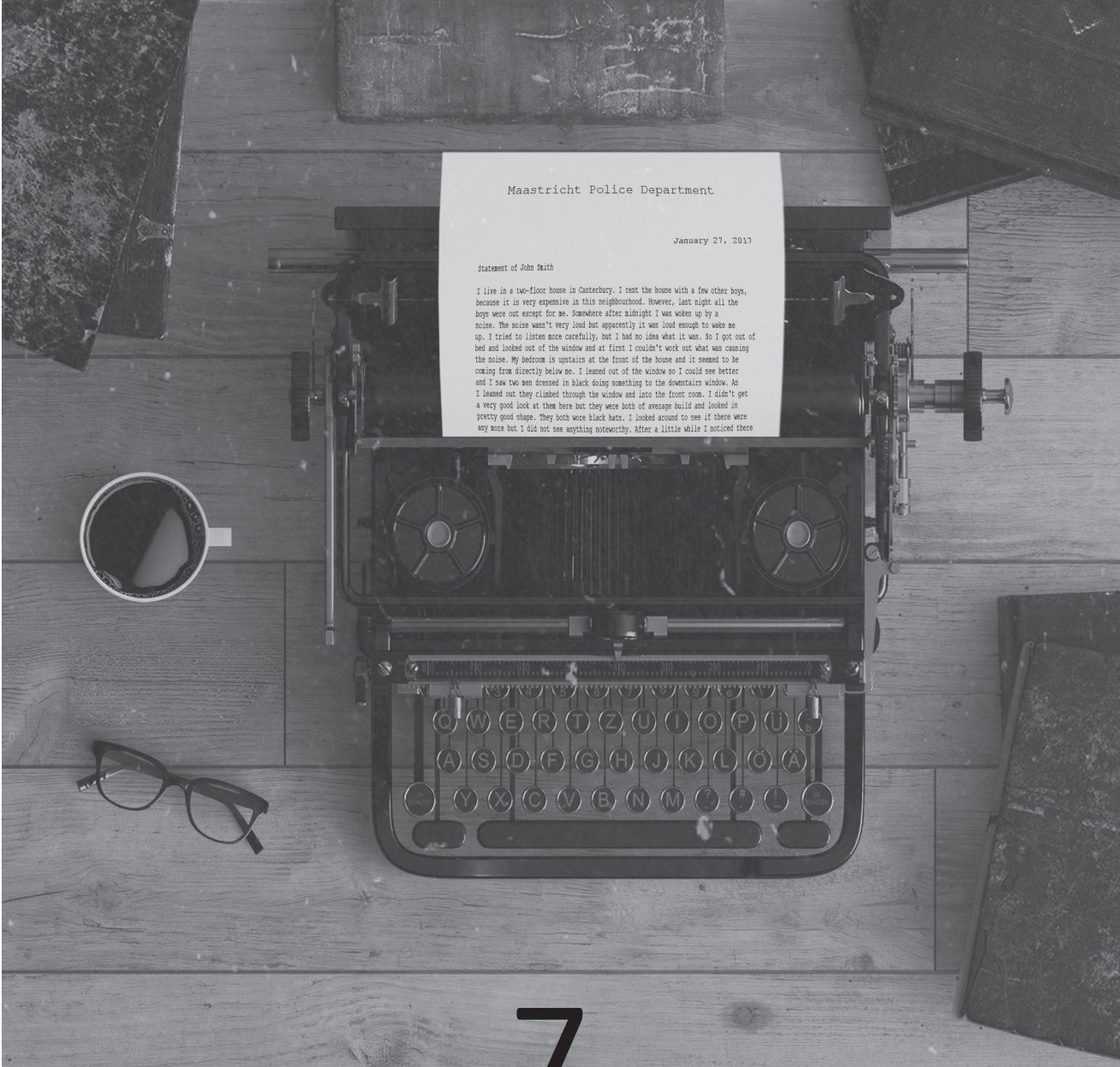
Reality Monitoring (RM)

1. *Clarity*: Refers to the clarity and vividness of the statement.
2. *Perceptual information*: Refers to the presence of sensory information in a statement (sound, smells, tastes, physical sensations and visual details.
3. *Spatial information*: Refers to information about locations or the spatial arrangement of people and/or objects.
4. *Temporal information*: Refers to information about when the event happened or explicitly describes a sequence of events.

5. *Affect*: Refers to information that describes how the participant felt during the event.
6. *Reconstructability of the story*: Examines whether it is possible to reconstruct the event on the basis of the information given.
7. *Realism*: Examines whether the story is plausible, realistic and makes sense.
8. *Cognitive operations*: Refers to descriptions of inferences made by the participant ant the time of the event. *

Scientific Content Analysis (SCAN)

1. *Social introduction*: Refers to how the persons described in the statement are introduced.
2. *Structure of the statement*: Refers to the balance of the statement. In a truthful statement 20% is used to describe activities leading up to the event, the next 50% to describe the actual event, and the final 30% to discuss what happened after the event.
3. *Emotions*: Refers to whether there are emotions described in the statement.
4. *Objective and subjective time*: Refers to how different time periods are covered in the statement. Objective time refers to the actual duration of events described, whereas subjective time refers to the amount of words spent to describe these events.
5. *First person singular, past tense*: Refers to the format in which a statement is written. Deviations from first person singular, past tense can indicate deception.
6. *Pronouns*: Refers to the use of pronouns in the statement. Omitting pronouns suggests reluctance of the writer to commit him/herself to the described actions.
7. *Spontaneous corrections*: Refers to all corrections that are made within the statements. *
8. *Lack of conviction or memory*: Refers to when the writer is vague about certain elements within the statement (e.g. “I think…”, “I guess…”) or when the writer admits he or she has forgotten something. *
9. *Out of sequence and extraneous information*: Examines whether the statement recounts the events in chronological order and whether there is extraneous information that does not seem relevant. *
10. *Missing information*: Refers to phrases in the statement that indicate that some information has been left out (e.g., finally, later on). *
11. *Change in language*: Refers to the change of terminology or vocabulary in the statement. A change in language indicates that something has altered in the mind of the writer. *



7

General Discussion

AIMS OF THIS DISSERTATION

This dissertation evaluated the use of three verbal credibility assessment methods, namely Scientific Content Analysis (SCAN), Criteria Based Content Analysis (CBCA), and Reality Monitoring (RM). The aim of this dissertation was twofold: (1) evaluate the usefulness of SCAN as a lie detection method and (2) investigating boundary conditions and possible improvements for all three methods. In doing so, we could draw conclusions about SCANs contribution to deception detection, and possible adaptations of CBCA and RM to improve their discriminating potential.

CAN SCAN BE USED AS A LIE DETECTION TOOL?

Part 1 primarily focused on investigating SCAN and its properties. Chapter 2 examined beliefs about verbal and nonverbal cues to deception and how these beliefs are related to the popularity of SCAN. However, participants did not strongly endorse the criteria, so SCAN's popularity cannot be explained by its intuitively appealing items. Furthermore, given that the actual number of SCAN criteria is unknown, Chapter 3 investigated which SCAN criteria were most often applied to real cases. To this end, we examined the SCAN analyses of 82 sexual abuse cases. We found that SCAN was primarily driven by a set of 12 criteria, very similar to those already published by Vrij (2008b). These 12 criteria served as a basis for our subsequent studies investigating SCAN. Moreover, all 82 cases were analyses by at least two SCAN analysts, which allowed us to study the inter-rater agreement of SCAN. SCAN's reliability showed to be unacceptably low, with an average reliability of 31% (Bogaard et al., 2014b: see Chapter 3).

Note that this is not the case for the studies presented in Chapters 4 to 6. In these studies, we resolved the problem of low inter-rater reliability by providing participants with a fixed list of SCAN criteria and their definitions, and a fixed coding scheme. In this way, we could investigate SCAN's ability to discriminate between truthful and fabricated statements. See Table 7.1 for an overview of SCAN's inter-rater reliability in the current thesis.

Table 7.1. Overview of SCAN's inter-rater reliability in the current thesis.

Chapter	Study	N	Which agreement	Inter-rater agreement	Type
3	Field	82	Criteria	0.31	Proportion agreement
4	Lab	117	Coding scheme	0.77	Cohen's Kappa
5	Lab	128	Coding scheme	0.79	Within-group interrater agreement (R_{wg})
6	Lab	64	Coding scheme	0.72	Cohen's Kappa

This low reliability is especially striking given all SCAN analysts followed Sapir's three-day SCAN training, and some of them also followed the *advanced* SCAN training (an additional two-day SCAN training). Moreover, all four analysts employed the same definitions - given in a self-produced summary - and should therefore have been able to analyse these criteria and their appropriate interpretations in a similar way. Nonetheless, SCAN proponents attribute this low reliability findings to the variability in SCAN expertise; had we made use of SCAN analysts with sufficient training and experience, we would certainly have found other results (Bockstaele, 2015). Whether this reasoning is correct, is highly debatable. After finishing the *basic* three-day SCAN course, trainees receive an official certificate, which should enable them to apply SCAN in a reliable way. Claiming that the SCAN training is insufficient, raises the question of when a SCAN analyst is sufficiently trained to apply the method. If this is not clearly defined, it can lead to circular reasoning: SCAN is a reliable tool, because if empirical evidence shows it not to be, this is due to the insufficient experience and quality of the SCAN analyst, not the tool (Bogaard, Meijer, & Merckelbach, 2016).

As Chapter 3 clarified which criteria are most often used when applying SCAN, Chapter 4 aimed at investigating whether these 12 criteria can discriminate between truthful and fabricated statements. To test the SCAN criteria, we used a standardized scoring system to rate the presence or absence of these criteria within statements. More precisely, all criteria were scored on a 3-point scale, ranging from 0 (not present) to 2 (strongly present). It is important to note that this approach deviates from how SCAN is applied in practice. SCAN does not consist of a fixed set of criteria or a standardized scoring system, and –as explained above – this results in a low reliability. However, without sufficient reliability, the validity of the SCAN criteria cannot be investigated. As such, Chapter 4 should best be interpreted as an examination of the validity of the most commonly used SCAN criteria, instead of an investigation of the SCAN approach. Based on the findings of Chapter 4, it is concluded that the application of SCAN in its current form should be discouraged. It could be argued that we have not tested the SCAN approach per se, and that this conclusion is not supported by the data presented in the current dissertation. Yet, from our findings that (1) SCAN's interrater reliability in the field is too low and (2) SCAN criteria in the lab do not accurately discriminate between truth tellers, we believe we can logically infer that the less systematic SCAN approach – as it is applied in practice - will not fare any better.

Chapter 4 indicated that only one criterion (i.e., change in language) showed to be of possible interest for lie detection. More precisely, liars included significantly more changes in language (e.g., "the knife" when the weapon is first mentioned in statement to "the blade" the second time it is mentioned) than truth tellers. This is interesting for deception research, because these changes - as far as we know - have not been described by any other verbal lie detection method in the deception literature. Nonetheless, because of Sapir's vague descriptions of SCAN criteria, it is unclear how this criterion exactly indicates deception. For example, the SCAN manual reads that changes in

language indicate deception, unless the change can be "justified". What Sapir exactly means with "a justified change" is not further specified in the manual. As such, the current study included all changes in language and an indication of deception, without interpreting whether the change was justified. To be of any value, its definition should be more strictly redefined, together with its interpretation, which should be tested in future studies.

Furthermore, investigation of the complete set of criteria - by summing up the separate criteria scores to get a total SCAN score - showed that the investigated SCAN criteria were unable to accurately discriminate between truthful and fabricated statements. Relying on these criteria when detecting deceit resulted in a correct classification of 50% of the deceptive statements and 53% of truthful statements. Several explanations can be formulated for SCAN's chance level accuracy.

First, the SCAN approach does not consist of criteria that investigate overall quantity of details and sensory perceptions, even though research has provided sufficient evidence that these criteria work well when assessing credibility (Amado et al., 2015; Amado et al., 2016; Masip et al., 2005; Oberlader et al., 2016; Vrij, 2005, 2008b). In contrast to what some SCAN proponents might argue (Bockstaele, 2008a, 2008b), these criteria are not included in the SCAN manual and are not explained during the SCAN training. Second, interpretations of several SCAN criteria are opposite to the interpretations of RM and CBCA. As the latter two methods are better embedded within scientific theories, their interpretation should be favoured over SCAN's interpretation. Third, as pointed out in Chapter 3 (Bogaard et al., 2014b), SCAN users do not agree about the exact number of criteria that need to be checked when analysing statements. Moreover, some criteria are only rarely present within statements, which makes them less useful for credibility assessment, but they can create noise nonetheless. Last, SCAN users generally only interpret the individual criteria and do not apply SCAN as a whole. The latter is a challenge because of the uncertainty of the exact number of criteria. Nevertheless, scholars strongly discourage to draw conclusions based on individual criteria as a tool's reliability is a function of the number of valid items (Amado et al., 2015; Amado et al., 2016; Masip et al., 2005; Oberlader et al., 2016; Vrij, 2005, 2008b). The higher this number, the more reliable the instrument.

SCAN DOES NOT WORK

So far, our studies yielded no supporting evidence for SCAN as a lie detection method; Chapter 3 showed that the interrater-reliability in the field was very low (Bogaard et al., 2014b), and Chapter 4 showed that when relying on SCAN criteria, lie detection accuracy does not exceed chance level (Bogaard et al., 2016).

However, it can be argued that the lack of significant differences might be due to the design and/or type of statements that we used. More precisely, we employed a

within subject design and asked participants to describe one truthful and one fabricated autobiographical event. We aimed at making the ecological validity as high as possible in an experimental setting, but realize that these statements are still different from the types of statements verbal credibility assessment methods are designed for. Therefore, in Part 2, the accuracy of SCAN is compared with CBCA, and RM, and research has indicated these latter tools can discriminate between truthful and fabricated statements. Leaving out the experimental manipulations from Chapter 5 and 6, thus solely investigating the main effect of veracity, showed once more that the investigated SCAN criteria were unable to distinguish between the two types of statements. In contrast, truthful statements did score significantly higher than fabricated ones when analysed with CBCA and RM. In both chapters, participants received the instructions to come up with one truthful and one fabricated statement about a recent negative event, such as a traffic accident in which they were involved, or a story about their wallet that got stolen. These instructions were similar to the ones given to participants in Chapter 4, demonstrating that the lack of significant SCAN findings cannot be attributed to our design or the type of statements that were used.

Another factor that could influence SCAN's accuracy is reliance on contextual information when conducting a SCAN analysis. As mentioned in Chapter 5, CBCA, RM and SCAN differ on their guidelines about taking into consideration contextual information. More precisely, CBCA encourages the use of contextual information; RM has no specific guidelines concerning extra information, and SCAN explicitly forbids the use of any additional information other than the statement itself. To make conclusions about the usage of contextual information, Chapter 5 investigated to what extent the accuracy of these methods was influenced by taking contextual information into consideration, and the next section considers the value and limitations of contextual information in detecting deception.

THE VALUE OF CONTEXTUAL INFORMATION

Whether contextual information should be taken into consideration for content analysis is an outstanding issue for credibility assessment. Information about how vulnerable CBCA, RM, and SCAN are to contextual bias is relevant in light of guidelines concerning the handling of this extra-domain information. Therefore, Chapter 5 investigated to what extent the credibility assessments of CBCA, RM, and SCAN were influenced by contextual information. For all tools, it was shown that statements presented with positive context cues (e.g., portraying the interviewee as a trustworthy person) were judged to be richer in criteria than statements presented with negative context cues (e.g., portraying the interviewee as an untrustworthy person) (Bogaard et al., 2014a). This finding clearly demonstrates that contextual information influences subsequent credibility assessment.

One possible solution to this contextual bias could be that someone unfamiliar to the case would perform the CBCA analyses, thereby separating judgements on the credibility of the *witness* from judgments on the credibility of the *statement*. However, allowing no information at all, as is the case with SCAN, is rather extreme and can also have a detrimental effect on credibility assessment. For example, Steller (1989) stated that information on the cognitive and verbal competence of the interviewee is necessary for content analysis. Indeed, research has shown that age and verbal competence can influence CBCA and RM scores and should be taken into consideration when interpreting these scores (Ruby & Brigham, 1997; Vrij et al., 2002). Therefore, based on our findings, I encourage practitioners to exclude personal history/character of the interviewee (as we included in our study as contextual information) when analysing statements. This information does not directly influence the quality and content of the statement, but does influence the interpretation of the results by the rater. On the other hand, information about personal characteristics and the interviewing procedures by which the statement is obtained, should be taken into consideration as they are necessary to make an appropriate interpretation of the CBCA and RM scores (Vrij et al., 2004b; Vrij et al., 2000).

Contextual information can also be diagnostic. Research investigating how people detect lies outside of the laboratory showed people rely a lot on contextual information when detecting lies, such as contradictions given by third parties (Park, Levine, McCornack, Morrison, & Ferrara, 2002). Indeed, using contextual information can increase our lie detection accuracy (Blair, Levine, & Shaw, 2010; Bond Jr, Howard, & Hutchison, 2013; Masip & Herrero, 2015). Blair et al. (2010) reported an average accuracy of 75% when taking into consideration contextual information, which is higher than the average accuracy reported for RM or CBCA. In their studies, contextual information was presented as either evidence-statement contradictions or information on base-rates of deception for the particular study. Bond Jr et al. (2013) provided participants with incentives to either lie or tell the truth. Observers were asked to make judgments about the veracity of the provided accounts. When participants were unaware about the incentives and purely based their judgments on behavior, average accuracy was 57%. In contrast, when observers were informed about the incentives, average accuracy increased up to 97%. It is important to note, however, that all participants gave an account following the incentive provided by the experimenter. The authors therefore concluded that these incentives were diagnostic contextual cues and should be taken into consideration when detecting deception. How these results translate into practice is unclear. In police investigations, guilty suspects all have an incentive to lie. Nevertheless, the difficulty lies in discriminating between guilty and innocent suspects. How these incentives can help in the discrimination process is not described.

Additionally, it is important to draw a distinction between contextual information for fact-finding and for lie detection purposes. Recently, Blair and colleagues (2010) have proposed three types of contextual information: (1) contradictions, (2) normative in-

formation or (3) idiosyncratic information. The first type refers to evidence that explicitly contradicts information that is provided by the interviewee (e.g., interviewee states that he was at home while video footage shows he was at a bus stop). Normative information refers to information about what is normal or possible in certain situations (e.g., suspect states she drove somewhere under 10 minutes while this normally takes 20 minutes, which means this person is probably lying). Idiosyncratic information is comparable to circumstantial evidence. For example, money shortage at work stops when a specific employee goes on vacation and begins again after he returns, making it highly likely this employee is involved.

Thus, contextual information is meaningful, and is especially important during the first stages of police investigation. Even though information is checked during this phase, one can debate whether this constitutes lie detection. One can argue that lie detection refers to a process where observable cues (i.e., visual, vocal or verbal) of the statement or interviewee are taken into consideration to form a judgment about the truthfulness of the statement. As such, Blair et al.'s notion of contextual information is related to fact-finding, not lie detection. Given that lie detection is notoriously difficult (Aamodt & Custer, 2006; C. F. Bond & DePaulo, 2006, 2008), and the error margins for credibility assessment methods are high (Amado et al., 2015; Amado et al., 2016; Masip et al., 2005; Oberlader et al., 2016; Vrij, 2005) lie detection should be seen as a last resort for testing the credibility of statements.

Aside from testing the weaknesses of verbal credibility assessment, such as contextual bias, this dissertation also focused on improving CBCA, RM, and SCAN's accuracy. Given that the error margins for CBCA and RM are high and should be decreased, Chapter 6 aimed at investigating possible manipulations to increase the accuracy of CBCA, RM and SCAN, which will be discussed in the next section.

IMPROVING THE ACCURACY OF CREDIBILITY ASSESSMENT METHODS

Chapter 6 focused on improving the accuracy of CBCA, RM, and SCAN by supplying individuals with a written example of a detailed statement. Although supplying participants with a detailed example did result in longer statements and higher CBCA, RM, and SCAN scores, it did so for both fabricated and truthful statements, failing to increase their accuracy (Bogaard, Meijer, & Vrij, 2014). A recent study by Leal, Vrij, Warmelink, Vernham, and Fisher (2015) also investigated the use of a model statement. In contrast to the written model statement in our study, they presented participants with a detailed audiotaped statement. Results were comparable to ours and showed that a model statement increased the number of words, but not CBCA's accuracy. The authors further investigated plausibility of the statements and concluded that truth tellers' stories sounded more plausible than liars' stories, but only in the model condition. Thus, providing individuals with a model statement elicits specific cues to detection that were not tapped by CBCA.

Another manipulation that has been used to increase the CBCA and RM's accuracy is the timing and amount of evidence that is disclosed to the interviewee. McDougall and Bull (2014) let participants take part in a mock crime. Liars were asked to go to a professor's office and copy exam answers, while truth tellers were asked to go to the same office but only copy an email address. After their return to the lab, the experimenters either began the interview with the disclosure of all the available evidence about their whereabouts (e.g., eyewitness saw them in the office) or gradually released this evidence throughout the interview. Next, all statements were transcribed and analysed with CBCA and RM. As expected, both CBCA and RM scores were higher for truthful accounts than for fabricated ones. However, when evidence disclosure was taken into consideration, only RM scores were significantly different for liars and truth tellers in the early disclosure condition. With gradual disclosure, no significant differences emerged between truthful and fabricated statements. So, gradual disclosure did increase the number of details liars included in their statements, but did not increase the accuracy of CBCA and RM. This is similar to our findings for the example statement. However, both interviewing styles led to an increased statement-evidence inconsistency for liars compared with truth tellers.

The finding that both the model statement and release of evidence did not increase CBCA and RM's accuracy, but generated other cues, indicates that these manipulations could be used in combination with other lie detection tools. One example could be the "verifiability approach" proposed by Nahari, Vrij, and Fisher (2014a). According to Nahari et al. (2014a) coming across as a truthful person poses a dilemma for liars. On the one hand, they want to give a statement that is rich in details to avoid suspicion, but on the other hand, they must be careful which details they include in their story to avoid being caught. To resolve this conflict, liars are expected to include less verifiable details. For example, if you testify you were withdrawing money at a specific bank, this is verifiable. However, if you testify you saw a red Volvo parked in a street without CCTV, this is not verifiable. So far, research investigating this approach shows promising results.

Several studies reported that liars indeed include less verifiable details in their statement compared with truth tellers, and that the ratio verifiable/unverifiable details is higher for truth tellers. Results concerning the differences in unverifiable details are mixed. Some studies report a higher number of unverifiable details for liars, while some report no difference (Nahari & Vrij, 2014; Nahari et al., 2014a; Vrij, Nahari, Isitt, & Leal, 2016). Interestingly, informing interviewees that their details will be checked caused truth tellers, but not liars, to include even more verifiable details, thereby actually facilitating lie detection (Nahari, Vrij, & Fisher, 2014b). Nahari et al. (2014b) instructed liars to commit a mock theft and to deny any involvement in this crime, while truth tellers were asked to give a truthful account of their activities. Prior to giving their statement, half of the liars, and half of the truth tellers were informed about the verifiability approach. They were told that the experimenter would check the details in their story,

they received information about verifiable and non-verifiable details, and were told that truth tellers are more inclined to include verifiable details in their stories. The authors labelled this the *information protocol*. Consistent with their expectations, informed truth tellers, but not informed liars, included more verifiable details in their stories. Similar findings have also been reported for the verifiability approach in insurance claims settings. Notifying claimants that their statements will be checked for verifiable details resulted in an overall accuracy rate of 80%, instead of 47.5% without the information protocol (Harvey, Vrij, Nahari, & Ludwig, 2016).

All in all, these manipulations (model statement and evidence) suggest that even though attempts to increase the accuracy of CBCA and RM have been little successful, they have led to alternative cues (e.g., inconsistencies and verifiability of details), which seem to be a valuable addition for lie detection.

ALTERNATIVE APPLICATIONS OF CONTENT ANALYSIS: ACID, ARJS AND PBCAT

Although this dissertation focused on investigating CBCA, RM, and SCAN, these are not the only verbal credibility assessment tools. One alternative is the Assessment Criteria Indicative of Deception (ACID; Colwell, Hiscock-Anisman, Memon, Taylor, & Prewett, 2008). ACID is based on the different cognitive and interpersonal demands of deception versus truth telling. It combines research into investigative interviewing, impression management, and memory. It is also embedded within a framework that requires a specific interview (i.e., reality interview, RI) for producing the statements that can be analysed later (Colwell, Hiscock-Anisman, & Memon, 2002), which is a variation of the Cognitive Interview (see Chapter 6; Bogaard, Meijer, & Vrij, 2014). The RI is designed to increase cognitive demands and points the interviewer to impression management tactics, the rationale is that these techniques facilitate recall for honest respondents, but at the same time make deception more difficult and obvious.

In the first phase of the interview the interviewee gives a free recall of his/her statement, next, s/he is asked to perform multiple recall tasks (i.e., mnemonic strategies) to increase cognitive demands. The statements are then transcribed and analysed with the following ACID criteria; length of the interviewee's response, potential errors that were admitted and the external (i.e., visual, auditory), contextual (i.e., spatial and temporal), and internal (i.e., cognitive operations) details derived from the RM theory. Four of these five criteria are coded in both the free recall and the mnemonics statement, resulting in a total of nine criteria. The criteria pertaining to the different types of details are coded based on their frequency, not on a 3-point scale as is done with CBCA and RM.

Colwell et al. (2008) showed that true statements were longer, more detailed, and contained more admissions of mistakes. Furthermore, ACID could accurately classify about 80% of the truthful and fabricated statements, while untrained raters accurately

scored approximately 56% of the statements. Similar findings were reported by Colwell et al. (2009), Suckle-Nelson et al. (2010), and Colwell, Hiscock-Anisman, and Fede (2013) for inmates', students', and children's statements. More recently Colwell, James-Kangal, Hiscock-Anisman, and Phelan (2015) have reported that an 8-hour ACID training improved lie detection accuracy from 54% pre-training to a remarkably high percentage of 89% post-training. Similar results have also been obtained for statements that were translated from Arabic to English (Morgan III, Colwell, & Hazlett, 2007). Thus, the ACID framework seems to be a promising technique to elicit verbal cues to deception.

It was previously mentioned that gradually disclosing information about evidence had detrimental effects on the quality of the statement for RM and CBCA, and the mnemonic strategies, as employed with ACID, have also shown to be ineffective for CBCA (Köhnken et al., 1995). Given that ACID is based on the RM theory, it is likely that RM would benefit from the aforementioned mnemonic strategies, but results have been mixed (Bembibre & Higueras, 2011; Hernandez-Fernaud & Alonso-Quecuty, 1997).

Another alternative to analyse the content of statements is the Aberdeen Report Judgment Scales (ARJS; Sporer, Masip, & Cramer, 2014). These scales are based on a factor analysis of both RM and CBCA criteria and builds upon theories from autobiographical memory, attribution theory, and self-presentational strategies. The ARJS consists of 52 items (including all CBCA and RM criteria) that are clustered into 13 scales, and every item is scored on a 7-point scale based on its quality, not quantity. However, as this version is long and time consuming, the authors have tested a Shorter Training Version, which is called ARJS-STV and consists of 17 items. In contrast to RM and CBCA, raters are given information on the weight of the different criteria. The 17 criteria are divided in three categories going from great weight (category 1) to small weight (category 3).

Although the authors concluded that the use of this method significantly improved accuracy (55%) compared to the control group (51%), this accuracy is only just above chance level. It is important to note, however, that the authors investigated false denials/accusations instead of completely fabricated statements, as is usually done in deception research. For example, liars were asked to tell about the misbehaviour they conducted, but to explain that someone else was responsible for this misconduct (i.e., false accusation). This means that liars only lied in a small part of their statements, making deception detection very hard, thus explaining the low accuracy rates.

The Psychologically Based Credibility Assessment Tool (PBCAT; Evans, Michael, Meissner, & Brandon, 2012) is another example of a verbal lie detection tool. The PBCAT was designed with two main concerns in mind. First, the authors wanted to use cues from different theoretical frameworks, and second, the method had to be user-friendly. The authors included cues from CBCA, RM, and the cognitive complexity theory. Cues selected from CBCA and RM were chosen based on their strong empirical support and whether they were easy to apply. A complete PBCAT consists of a list of 11 cues, seven investigating truthfulness and four investigating deception. Five criteria are scored based on a three-point scale, while the remaining cues are rated on a Likert scale

ranging from -2 to 2. In contrast to ACID, PBCAT does not make any recommendations as to how the statements should best be obtained.

With minimal training, the overall accuracy for PBCAT was only slightly higher than for controls (60% vs. 54%); better results were reported for the truth accuracy (72% vs. 58%), compared with the lie accuracy (50% vs. 54%). Research with native and non-native English speakers showed no incremental value for using PBCAT (Evans & Michael, 2013). Moreover, participants were not instructed on how to interpret the results of the PBCAT, just to indicate whether the statement was truthful or not. As such, it is unclear how the PBCAT criteria have influenced their judgments. Additionally, no inter-rater reliabilities were calculated, so it is impossible to know whether the PBCAT was applied reliably. Although most studies investigating CBCA and RM concluded that the accuracy of all criteria combined is better than investigating criteria separately (Amado et al., 2015; Amado et al., 2016; Masip, Sporer, et al., 2005; Oberlader et al., 2016; Vrij, 2005), PBCAT does not require its users to calculate a total score. As such, reliability and interpretation of PBCAT items should be investigated before conclusions can be drawn about the applicability of PBCAT as a lie detection tool.

AUTOMATED SCORING

Most credibility assessment tools (i.e., CBCA, RM, ACID, ARJS) have guidelines on how criteria should be interpreted and scored, yet the scoring is dependent on a coder. Reliability for most items is good but the scoring is subjective. To increase objectivity when scoring, Pennebaker, Francis and Booth (2001) developed the Linguistic Inquiry and Word Count (LIWC), which is a computer based program for analysing statements. More precisely, the LIWC identifies and counts specific content and style words by comparing these words on a word-by-word level with a file consisting of more than 2000 words, divided in 72 linguistic dimensions. Thereby investigating quantity (i.e., frequency) instead of quality (e.g., CBCA and RM). The investigated dimensions include, for example, words describing negative and positive emotions (e.g., happy, sad), pronouns (e.g., you, mine, ours), and various content categories (i.e., profession, religion) (Chung & Pennebaker, 2007; Pennebaker, Francis, & Booth, 2001). The output is given in percentages of the total words that belong to a specific dimension. Thus, LIWC gives a summary of linguistic features in a written statement based upon the predefined categories (Newman et al., 2003).

Newman et al. (2003) indeed showed that the LIWC can differentiate between liars and truth tellers. Liars, compared with truth tellers, used fewer first and third person pronouns (e.g., I and he respectively), more words expressing a negative emotion, fewer exclusive words (e.g., except, but) and more motion verbs (e.g., swim, run). Moreover, Bond and Lee (2005) investigated the truthful and deceptive language of prisoners with the five categories mentioned above and an automated version of RM in which the

LIWC categories "sensory words", "spatial words", "temporal words", and "cognitive mechanism words" were included. Based on the five categories, LIWC could classify 69.7% of the statements correctly, while the RM correctly classified 71.1%. More recent, Driskell, Neuberger, Driskell, Burke, and Salas (2014) reported similar findings for automated RM.

In sum, even though verbal indicators were scored automatically, the accuracy rates reported for the LIWC models are similar to the accuracy rates of RM (Masip et al., 2005) and CBCA (Amado et al., 2015; Amado et al., 2016; Oberlader et al., 2016; Vrij, 2005). However, while verbal lie detection methods have primarily focused on the content of the writers' narratives, LIWC studies mainly focused on investigating writing style. Furthermore, until now, the LIWC program is unable to recognize expressions and proverbs, while human raters can. Consequently, it is likely that there is a trade-off between objectivity of automatic scoring and the inability to process content.

IDENTIFYING DECEPTION ONLINE

One of the most important applications of automated scoring systems is to detect deception online. Increasing numbers of words are being digitized and published online on social platforms such as Twitter and Facebook, but also via email, product reviews or dating websites, yet the veracity of this information is often questionable. Taking into consideration the ease with which these reviews can be posted, it is likely that groups of spammers are working together to influence the sentiment of websites or products. For example, they can try to boost the sale on a specific product by fabricating positive reviews, or by downgrading competing products. Ott, Cardie, and Hancock (2012) used an automated prediction model to estimate the prevalence of false reviews on Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor, and Yelp. They reported highest spam rates for TripAdvisor (i.e., 6%), and an average spam rate of around 2%-4% overall. Deceptive reviews are a growing problem, especially for websites with low requirements for posting reviews. The authors argued that increasing the requirements, for example by asking reviewers to write at least 2 reviews before they are published online, lowers spam prevalence.

The finding that the veracity of messages can be analysed by means of algorithms has inspired many scholars to develop algorithms tailored to these online systems. Wu, Greene, Smyth, and Cunningham (2010) investigated 30,000 TripAdvisor reviews for the proportion of positive singletons (i.e., positive reviews from first-time reviewers) and the concentration of positive singletons. These are clusters of positive reviews that occur within the same timeframe. Their algorithm assumes that the more temporal clustering between positive singletons, the more likely it is that these reviews are false. Based on these numbers, the authors could identify reviews that were suspicious. Next, the authors correlated the original ranking of the hotel with the ranking after removal of the suspicious reviews, and the higher this correlation, the more reliable the ranking of the hotel.

A more sophisticated approach has been tested by Lim, Nguyen, Jindal, and Lauw (2010), who investigated opinion spammers on Amazon. Their software is specifically aimed at identifying spammers by means of four models that take into consideration whether (1) reviews are identical or look similar, (2) the same user reviews many different products with the same attributes, (3) the time frame in which the reviews are written by the same user and (4) the general deviation of the specific review to the other reviews. Investigation of these properties leads to a total spam score that can be attributed to specific users. This algorithm showed to be valuable for identifying spammers. However, ground truth for the reviews remains unknown. To obtain an indication of ground truth labels for the users, the authors have tested previously established algorithms.

Social network sites, such as Twitter and Facebook, suffer from a spamming problem. Potential threats to these communities are cybercriminals who are interested in exploiting users by alluring them to malignant websites (e.g., illegal loans, insurances, products) or using their personal information for future spamming (i.e., identity theft). Many different algorithms have been used to successfully identify these spammers, with reported accuracies up to 95% (Benevenuto, Magno, Rodrigues, & Almeida, 2010; Chu, Widjaja, & Wang, 2012; Lee, Caverlee, & Webb, 2010; Stringhini, Kruegel, & Vigna, 2010). Based on these computer models, over 15,000 fake accounts have been deleted from Twitter (Chu et al., 2012).

As becomes clear, the reported accuracy rates of models testing deception online are higher than models testing human deception in daily life. One important reason for this discrepancy is that these algorithms consider a lot more information than only the content of the message. For example, the URL the message has been posted from, the account it has been sent from, users who are linked to this account, and which website the message is linked to, are all included in the model. Hence, in contrast to detecting deceit with credibility assessment tools, online deception detection can only be successful when algorithms consider contextual information.

HOTSPOTS

This dissertation has focused on investigating SCAN as a lie detection tool. However, next to detecting deceit, a second purpose of SCAN is to aid in generating further specific questions for police investigations (Sapir, 2005). According to Sapir, some SCAN criteria are included because they highlight “sensitive information” within a statement (e.g., unasked explanations; unimportant information), also called hotspots. It is indeed claimed that SCAN is useful during the early stages of an investigation because it aids in highlighting these hotspots (Smith, 2001) that can be used for generating questions for the subsequent interrogation.

However, peer-reviewed research investigating whether detectives generate more or better questions when relying on SCAN is lacking. To my knowledge, only one thesis has investigated this question, and will therefore be included here. In this thesis, Van Geest (2008) investigated whether (1) SCAN analysts detect more hotspots within a statement than detectives who do not use SCAN; (2) there is a difference in question content between SCAN analysts and detectives who do not use SCAN, and (3) SCAN analysts ask questions that are of better quality than detectives who do not rely on SCAN. Fifteen participants (two SCAN trained detectives) were asked to investigate two statements and to highlight the hotspots within these statements. A hotspot was defined as "words or phrases where you doubt the veracity and/or completeness". Next, participants were asked which question(s) they would formulate for each hotspot they marked. Outcomes showed no difference between the number of hotspots or the number of questions formulated with or without SCAN. Question content was investigated by categorizing them as who, what, where, when and why questions, but again no differences were reported between detectives who used SCAN and those who did not. Finally, question quality was tested by categorizing the questions as open versus closed and suggestive versus non-suggestive questions, but results exhibited no indication of a superior SCAN performance.

These SCAN hotspots are very similar to Ekman's clusters of clues that he claims to be indicative of deception (Ekman, 2009). Ekman stated that lie detectors should be looking for "hotspots", which are discrepancies between the story content and visual and vocal cues people display when telling the story, indicating something is off. Moreover, he claimed these hotspots allow high accuracy in distinguishing liars and truth tellers (Ekman, 2014). Although Ekman also included vocal cues as hotspots, most his research has focused primarily on behavioral cues and facial expressions (i.e., micro expressions). For these hotspots to be of any value, two claims should be empirically supported. First, lying should indeed result in behavioral and facial cues, and second, these cues should be detectable through observation.

The first claim is already problematic, as an extensive meta-analysis by DePaulo et al. (2003) showed that behavioral cues to deception are scarce, and if there is a relationship, the effect sizes are small. Furthermore, research about facial expressions showed that expressions do not differ between experienced and fabricated emotions (Porter & ten Brinke, 2008). The second claim assumes that micro expressions should be detectable through observation. However, Porter, ten Brinke, and Wallace (2012) concluded that differences in micro-expressions do exist between truth tellers' and liars' emotional faces, but they only exist as partial expressions in either the lower of upper half of the face. This makes them more difficult to notice than is claimed by Ekman. Moreover, these results are restricted to deceptive emotional expressions (e.g., displaying happiness when feeling disgust) and by no means are similar to behavior expressed when telling a fabricated story. Taking into consideration that most nonverbal cues are not diagnostic of deception and that micro expressions are difficult to observe; the

correctness of Ekman's claims is highly doubtful. Thus, the search for hotspots, as indicated by Sapir and Ekman, has not been proven helpful in detecting deceit.

FUTURE RESEARCH

Sapir's website and other websites teaching deception detection claim that lie detection is an easy task and becomes a "second nature" after training (retrieved from http://www.leugenacademie.nl/home.html and http://www.lsiscan.com). This dissertation clearly showed that lie detection is not that simple. Although researchers have invested time and effort in adapting and improving alternatives to the CBCA and RM framework, such as ACID, ARJS and PBCAT, error rates reported are still similar to those of CBCA and RM.

Chapter 5 demonstrated the problems that arise with the current framework CBCA is embedded in. Therefore, future studies should focus on developing better protocols for verbal lie detection. Research has already demonstrated that cues to deception are weak and unpredictable (Bond & DePaulo, 2006; DePaulo et al., 2003), so a first step for improving lie detection protocols could be to focus on interviewing techniques that may increase the elicitation of diagnostic cues. Previous studies demonstrated that lying is more cognitively demanding than telling the truth (see also Vrij et al., 2006; 2008) and interviewers could manipulate this difference by introducing tactics known to increase cognitive load. According to Vrij and Granhag (2012) there are two ways to achieve this: (1) by using tactics that increase the difficulty of recalling an event such as telling the story backwards, or (2) by using tactics to increase talkativeness, such as priming the interviewee with an example statement (see Chapter 6). Research has indeed shown that telling a story in reverse order increases lie detection accuracy (Vrij et al., 2012; Vrij et al., 2008). Another way of obtaining more information from a suspect is to use an information gathering approach instead of an accusatory interview style. The latter technique confronts the suspect with accusations, whilst the former uses a series of open-ended questions that have been shown to increase cognitive load, and encourage interviewees to offer more information (Vrij, Mann, & Fisher, 2006). Information is key when it comes to lie detection, as research has indicated content cues to be more useful for deception purposes than nonverbal cues (DePaulo et al., 2003).

Therefore, research should provide better guidelines into the specific open questions that should be asked during an interview. Research investigating how asking unanticipated questions influenced lie detection showed that pairs of liars had less overlap in their stories for unanticipated than anticipated questions (Vrij et al., 2009). More recently, Warmelink, Vrij, Mann, Jundi, and Granhag (2012) investigated liars and truth tellers' answers individually to anticipated and unanticipated questions about a planned activity. Participants were interviewed about a trip they were planning to make in the future, and half of the participants were instructed to lie about their trip. Anticipated

questions pertained to the main purpose of their trip, while the unanticipated questions were about transportation and planning issues (e.g., What part of the trip was most difficult to plan?). Liars provided more details to the anticipated questions, due to a better preparation, but fewer details to the unanticipated questions. Especially the ratio in the number of details generated by both types of questions (i.e., number of details anticipated / number of details unanticipated) seems to be very promising in discriminating truth tellers and liars. Similar results were reported by Sooniste, Granhag, Knieps, and Vrij (2013).

It might also be interesting to investigate how a combination of human and automatic coding of statements could improve diagnostic accuracy. Chapter 5 showed that people incorporate irrelevant information from their credibility analysis, thus leading to different evaluations for identical statements. Moreover, human lie detectors rely on cognitive heuristics, overestimate dispositional factors (e.g., personality), and tend to focus too much on behavioral signs to detect deception instead of listening to the speaker (Levine & McCornack, 2001; Vrij, 2008c). Automatic coding, such as the LIWC, on the other hand, has the disadvantage it cannot (yet) understand the meaning of words (Pennebaker et al., 2001; Zijlstra, Middendorp, Meerveld, & Geenen, 2005), which could result in a wrongful word categorization for some text fragments. However, a recent meta-analysis combined the findings of 30 studies investigating linguistic markers by means of computer programs (Hauch et al., 2014) and showed that liars used fewer words, shorter sentences, and their stories were less complex compared to truth tellers. Additionally, liars used more negative emotion words, especially words describing anger, and distanced themselves more (i.e., less self-references and more other-references) from the fabricated events than truth tellers. Contrary to the RM theory, truth tellers used more words indicating cognitive operations compared with liars. Although these results seem encouraging, effect sizes were small (ranging from 0.11 to 0.20), probably because automated programs do not investigate the actual content. Perhaps a combination of both coding types could have an additional value for lie detection purposes.

In sum, to improve verbal lie detection, we should not underestimate the importance of how the statements are obtained. Shifting the focus from a passive way of acquiring statements to a more deliberate manipulation of cognitive load seems to show potential for refining verbal lie detection. Further research into the boundary conditions and practical applicability of these tactics might lead to the development of a theoretically based protocol of verbal lie detection. Several authors have suggested the need for a paradigm change and to step away from cue-based lie detection towards systems that primarily rely on between-statement and statement-evidence consistency when making deception judgments (Blair, Levine, Reimer, & McCluskey, 2012; Levine & McCornack, 2014). Still, this dissertation shows that verbal cues to deception are helpful. The proposed consistency system is closely related to the point I have raised about contextual information; lie detection should only be considered when all other path-

ways have been investigated, meaning that checking whether there is correspondence between a statement and the factual evidence is fact-finding and should be done before turning to lie detection methods.

Furthermore, when investigating the accuracy of verbal credibility assessment, researchers - the author of this thesis included - mainly focus on testing how well these approaches can discriminate between truthful and fabricated statements. However, according to the Signal Detection Theory (Swets, 2014) this is only the first step in testing diagnostic accuracy. Also important is the predictive power (positive and negative) of these approaches, and these depend on the base rate of fabricated statements. Laboratory research typically uses a base rate of 50%, which likely deviates substantially from the actual base rate of fabricated statements in police investigations (e.g., false allegations, false alibis). How important these base rates are, can be illustrated using the following example. Lisak et al. (2010) estimated that approximately 5% of all sexual allegations are false. Furthermore, research has shown that the average accuracy for CBCA and RM varies around 70%. Given 100 statements; CBCA and RM would classify 28 truthful statements as false (30% of 95) and 3 of the false statements correctly as false (70% of 5). Given the low base rate, this means that when a statement is classified as deceptive by applying CBCA or RM, this classification is wrong in 90% of the cases [28/(3+28)] (for a similar example see Merckelbach, 2015). If we apply this calculation to the SCAN method with a 52% accuracy (See Chapter 4) there is a 94% chance that a deceptive classification is wrong. This illustrates why practitioners should be very careful when applying credibility assessment tools, and that scholars should work on further refining these approaches.

In contrast to CBCA and RM, Sapir claims that his SCAN can generate categorical judgments with high precision. For example, he writes on his website that "SCAN will solve every case… and will show you whether the subject is truthful or deceptive" (retrieved from http://www.lsiscan.com/intro_to_scan.htm). Given there are no unique features that designate deception (DePaulo et al., 2003), the likelihood that the scientific field will ever be able to produce a method with hit rates over 95% is highly doubtful. Therefore, all methods that claim deception can be detected easily with high accuracy, such as SCAN, should be scrutinized and approached with great caution. Nonetheless, scholars do succeed in improving the accuracy of existing methods and testing new manipulations and boundary conditions, showing progression occurs gradually, yet steadily. Lie detection will always be an important issue in the legal field and advances in deception research should be highly encouraged.

Reference list

Aamodt, M. G., & Custer, H. (2006). Who can best catch a liar? A meta-analysis of individual differences in detecting deception. *Forensic Examiner, 15*, 6-11.

Adams, S. H. (1996). Statement analysis: What do suspects' words really reveal? *FBI Law Enforcement Bulletin*, 12-20.

Akehurst, L., Köhnken, G., & Höfer, E. (2001). Content credibility of accounts derived from live and video presentation. *Legal and Criminological Psychology, 6*, 65-83.

Akehurst, L., Köhnken, G., Vrij, A., & Bull, R. (1996). Lay persons' and police officers' beliefs regarding deceptive behaviour. *Applied Cognitive Psychology, 10*, 461-471.

Amado, B. G., Arce, R., & Fariña, F. (2015). Undeutsch hypothesis and Criteria Based Content Analysis: A meta-analytic review. *The European Journal of Psychology Applied to Legal Context, 7*, 1-10.

Amado, B. G., Arce, R., Fariña, F., & Vilariño, M. (2016). Criteria-Based Content Analysis (CBCA) reality criteria in adults: A meta-analytic review. *International Journal of Clinical and Health Psychology, 16*, 201-210.

Anson, D. A., Golding, S. L., & Gully, K. J. (1993). Child sexual abuse allegations: Reliability of criteria based content analysis. *Law and Human Behavior, 17*(331-341).

Armistead, T. W. (2011). Detecting deception in written statements: The British Home Office study of scientific content analysis (SCAN). *Policing: An International Journal of Police Strategies & Management, 34*, 588-605.

Arntzen, F. (1970). *Psychologie der Zeugenaussage*. Göttingen, Germany: Hogrefe.

Ask, K., Rebelius, A., & Granhag, P. A. (2008). The 'elasticity' of criminal evidence: A moderator of investigator bias. *Applied Cognitive Psychology, 22*, 1245-1259.

Bembibre, J., & Higueras, L. (2011). Differential effectiveness of the Cognitive Interview in a simulation of testimony. *Psychology, Crime & Law, 17*, 473-489.

Ben-Shakhar, G., Bar-Hillel, M., Bilu, Y., & Shefler, G. (1998). Seek an you shall find: A confirmation bias in clinical judgment. *Journal of Behavioral Decision Making, 11*, 235-249.

Benevenuto, F., Magno, G., Rodrigues, T., & Almeida, V. (2010). Detecting spammers on twitter. *Collaboration, electronic messaging, anti-abuse and spam conference, 6*, 12.

Blair, J. P., Levine, T. R., Reimer, T. O., & McCluskey, J. D. (2012). The gap between reality and research: Another look at detecting deception in field settings. *Policing: An International Journal of Police Strategies & Management, 35*, 723-740.

Blair, J. P., Levine, T. R., & Shaw, A. S. (2010). Content in context improves deception detection accuracy. *Human Communication Research, 36*, 423-442.

Blandon-Gitlin, I., Pezdek, K., Lindsay, S., & Hagen, L. (2009). Criteria-based Content Analysis of true and suggested accounts of events. *Applied Cognitive Psychology, 23*, 901-917.

Bockstaele, M. (2008a). De SCAN als middel tot waarheidsvinding [SCAN as a tool for searching the truth]. *Het Tijdschrift voor de Politie [Journal of Police], 70*, 8-13.

Bockstaele, M. (2008b). Scientific Content Analysis (SCAN). Een nuttig instrument bij verhoren? [SCAN: A valuable tool for interrogations?]. In L. Smets & A. Vrij (Eds.), *Het analyseren van de geloofwaardigheid van verhoren: Het gebruik van leugendetectiemethoden [The analysis of the credibility of interrogations: The use of lie detection methods]* (pp. 105-156). Brussels, Belgium: Politeia.

Bockstaele, M. (2015). Vragen bij de kwaliteit van rechtspsychologisch onderzoek over verhoortechnieken. *Panopticon, 36*, 375-384.

Bogaard, G., Meijer, E., & Merckelbach, H. (2016). Screenen met SCAN? Liever niet. *Panopticon, 37*, 197-210.

Bogaard, G., Meijer, E., & Vrij, A. (2014). Using an example statement increases information but does not increase accuracy of CBCA, RM, and SCAN. *Journal of Investigative Psychology and Offender Profiling, 11*, 151-163. doi:10.1002/jip.1409

Bogaard, G., Meijer, E., Vrij, A., Broers, N. J., & Merckelbach, H. (2014a). Contextual Bias in Verbal Credibility Assessment: Criteria Based Content Analysis (CBCA), Reality Monitoring (RM), and Scientific Content Analysis (SCAN). *Applied Cognitive Psychology, 28*, 79–90. doi:10.1002/acp.2959

Bogaard, G., Meijer, E., Vrij, A., Broers, N. J., & Merckelbach, H. (2014b). SCAN is largely driven by 12 criteria: Results from field data. *Psychology, Crime and Law, 20*, 430-449. doi:10.1080/1068316X.2013.793338

Bogaard, G., Meijer, E., Vrij, A., & Merckelbach, H. (2016). Scientific Content Analysis (SCAN) cannot distinguish between truthful and fabricated accounts of a negative Event. *Frontiers in Psychology, 7*, 1-7. doi:10.3389/fpsyg.2016.00243

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Individual Differences, 10*, 214-234.

Bond, C. F., & DePaulo, B. M. (2008). Individual differences in judging deception: Accuracy and bias. *Psychological Bulletin, 134*, 477-492.

Bond, G. D., & Lee, A. Y. (2005). Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language. *Applied Cognitive Psychology, 19*, 313-329.

Bond Jr, C. F., Howard, A. R., & Hutchison, J. L., & Masip, J. (2013). Overlooking the obvious: Incentives to lie. *Basic and Applied Social Psychology, 35*, 212-221.

Burgoon, J. K., Blair, P. J., & Strom, R. E. (2008). Cognitive biases and nonverbal cue availability in detecting deception. *Human Communication Research, 34*, 552-599.

Burton, P., Gurrin, L., & Sly, P. (1998). Tutorial in biostatistics. Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modeling. *Statistics in medicine, 17*, 1261-1291.

Caso, L., Vrij, A., Mann, S., & De Leo, G. (2006). Deceptive responses: The impact of verbal and non-verbal countermeasures. *Legal and Criminological Psychology, 11*, 99-111.

Chu, Z., Widjaja, I., & Wang, H. (2012). Detecting social spam campaigns on twitter. *Applied Cryptography and Network Security*, 455-472.

Chung, C., & Pennebaker, J. W. (2007). The psychological functions of function words. In K. Fiedler (Ed.), *Social Communication* (pp. 343-359). New York: Psychology Press.

Colwell, K., Hiscock-Anisman, C., & Fede, J. (2013). Assessment Criteria Indicative of Deception: An example of the new paradigm of differential recall enhancement *Applied issues in investigative interviewing, eyewitness memory, and credibility assessment* (pp. 259-291). New York: Springer.

Colwell, K., Hiscock-Anisman, C. K., & Memon, A. (2002). Interviewing techniques and the assessment of statement credibility. *Applied Cognitive Psychology, 16*, 287-300.

Colwell, K., Hiscock-Anisman, C. K., Memon, A., Colwell, L. H., Taylor, L., & Woods, D. (2009). Training in Assessment Criteria Indivative of Deception to improve credibilty judgments. *Journal of Forensic Psychology Practice, 9*, 199-207.

Colwell, K., Hiscock-Anisman, C. K., Memon, A., Taylor, L., & Prewett, J. (2008). Assessment Criteria Indicative of Deception (ACID): an integrated system of investigative interviewing and detecting deception. *Journal of Investigative Psychology and Offender Profiling, 4*, 167-180.

Colwell, K., James-Kangal, N., Hiscock-Anisman, C., & Phelan, V. (2015). Should Police Use ACID? Training and Credibility Assessment Using Transcripts Versus Recordings. *Journal of Forensic Psychology Practice, 15*, 226-247.

Curtis, D. A. (2015). Patient deception: nursing professionals' beliefs and attitudes. *Nurse Educator, 40*, 254–257. doi:doi:10.1097/NNE.0000000000000157

Curtis, D. A., & Hart, C. L. (2015). Pinocchio's nose in therapy: Therapists' beliefs and attitudes toward client deception. *International Journal for the Advancement of Counselling, 37*, 279-292.

Davis, M., Markus, K. A., & Walters, S. B. (2006). Judging the credibility of criminal suspect statements: Does mode of presentation matter? *Journal of Nonverbal Behavior, 30*, 181-198.

DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., & Epstein, J. A. (1996). Lying in everyday life. *Journal of Personality and Social Psychology, 70*, 979-995.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin, 129*, 74-118.

Driscoll, L. (1994). A validity assessment of written statements from suspects in criminal investigations using the SCAN technique. *Police Studies, 4*, 77-88.

Driskell, J. E. (2012). Effectiveness of deception detection training: A meta-analysis. *Psychology, Crime & Law, 18*, 713-731.

Driskell, T., Neuberger, L., Driskell, J. E., Burke, C. S., & Salas, E. (2014). The Language of Lies A Content Analytic Approach. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 58*, 1328-1331.

Dror, I. E., Charlton, D., & Péron, A. E. (2006). Contextual information renders experts vulnerable to making erroneous identifications. *Forensic Science International, 156*, 74-78.

Ekman, P. (2009). *Telling lies: Clues to deceit in the marketplace, politics, and marriage*. New York - London: W. W. Norton & Company.

Ekman, P. (2014). Catching Liars [retrieved from http://www.huffingtonpost.com/paul-ekman/catching-liars_b_5676176.html].

Elaad, E. (2003). Effect of feedback on the overestimated capacity to detect lies and the underestimated ability to tell lies. *Applied Cognitive Psychology, 17*, 349-363.

Evans, J. R., & Michael, S. W. (2013). Detecting deception in non-native English speakers. *Applied Cognitive Psychology*.

Evans, J. R., Michael, S. W., Meissner, C. A., & Brandon, S. E. (2012). Validating a new assessment method for deception detection: Introducting a Psychologically Based Credibility Assessment Tool. *Journal of Applied Research in Memory and Cognition*, 33-41.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology, 43*, 543-549.

Findley, K. A., & Scott, M. S. (2006). The multiple dimensions of tunnel vision in criminal cases. *Wisconsin Law Review, 2*, 291-397.

Fisher, R. P., Amador, M., & Geiselman, R. E. (1989). Field test of the cognitive interview: Enhancing the recollection of actual victims and witnesses of crime. *Journal of Applied Psychology, 74*, 722-727.

Geiselman, R. E., Fisher, R. P., MacKinnon, D. P., & Holland, H. L. (1986). Enhancement of eyewitness memory with the cognitive interview. *American Journal of Psychology, 99*, 385-401.

Gödert, G. W., Gamer, M., Rill, H., & Vossel, G. (2005). Statement validity assessment: Inter-rater reliability of criteria-based content analysis in the mock-crime paradigm. *Legal and Criminological Psychology, 10*, 225-245.

Granhag, P. A., Andersson, L. O., & Strömwall, L. A. (2004). Imprisoned knowledge: Criminals' beliefs about deception. *Legal and Criminological Psychology, 9*, 103-119.

Granhag, P. A., Strömwall, L. A., & Hartwig, M. (2007). The SUE-technique: The way to interview to detect deception. *Forensic Update, 88*, 25-29.

Greer, E. (2000). The truth behind legal dominance feminism's "two percent false rape claim" figure. *Loyola Of Los Angeles Law Review, 33*, 947-972.

Gudjonsson, G. H., Sigurdsson, J. F., Asgeirsdottir, B. B., & Sigfusdottir, I. D. (2007). Custodial interrogation: What are the background factors associated with claims of false confession to police? *Journal of Forensic Psychiatry & Psychology, 18*, 266-275.

Halevy, R., Shalvi, S., & Verschuere, B. (2013). Being honest about dishonesty: Correlating self-reports and actual lying. *Human Communication Research, 40*, 54–72. doi:10.1111/hcre.12019

Hart, C. L., Hudson, L. P., Fillmore, D. G., & Griffith, J. D. . (2006). Managerial Beliefs about the Behavioral Cues of Deception. *Individual Differences Research, 4*, 176-184.

Hartwig, M., & Bond, C. F. (2011). Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychological Bulletin, 137*, 643-659.

Hartwig, M., & Bond, C. F. J. (2014). Lie detection from multiple cues: A meta-analysis. *Applied Cognitive Psychology*. doi:10.1002/acp.3052

Harvey, A. C., Vrij, A., Nahari, G., & Ludwig, K. (2016). Applying the Verifiability Approach to insurance claims settings: Exploring the effect of the information protocol. *Legal and Criminological Psychology*.

Hauch, V., Blandón-Gitlin, I., Masip, J., & Sporer, S. L. (2014). Are Computers Effective Lie Detectors? A Meta-Analysis of Linguistic Cues to Deception. *Personality and Social Psychology Review*, 1-36.

Hauch, V., Sporer, S. L., Michael, S. W., & Meissner, C. A. (2014). Does training improve the detection of deception? A meta-analysis. *Communication Research*, 1-61.

Hernandez-Fernaud, E., & Alonso-Quecuty, M. L. (1997). The Cognitive Interview and lie detection: A new magnifying glass for Sherlock Holmes. *Applied Cognitive Psychology, 11*, 55-68.

Hill, C., Memon, A., & McGeorge, P. (2008). The role of confirmation bias in suspect interviews: A systematic evaluation. *Legal and Criminological Psychology, 13*, 357-371.

Horowitz, S. W., Lamb, M. E., Esplin, P. W., Boychuk, T. D., Krispin, O., & Reiter-Lavery, L. (1997). Reliability of criteria-based content analysis of child witness statements. *Legal and Criminological Psychology, 2*, 11-21.

Iacono, W. G. (1991). Can we determine the accuracy of polygraph tests? In J. R. Jennings, P. K. Ackles, & M. G. H. Coles (Eds.), *Advances in psychophysiology* (Vol. 4, pp. 201-207). London: Jessica Kingsley Publishers.

James, L. R., Demaree, R. G., & Wolf, G. (1993). R_{wg}: An assessment of within-group interrater agreement. *Journal of Applied Psychology, 78*, 306-309.

Johnson, M. K., Foley, M. A., Suengas, A. G., & Raye, C. L. (1988). Phenomenal characteristics of memories for perceived and imagined autobiographical events. *Journal of Experimental Psychology: General, 12*, 371-376.

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review, 88*, 67-85.

Johnson, M. K., Raye, C. L., Foley, H. J., & Foley, M. A. (1981). Cognitive operations and decision bias in reality monitoring. *The American Journal of Psychology, 94*, 37-64.

Jones, M., & Sugden, R. (2001). Positive confirmation bias in the acquisition of information. *Theory and Decision, 50*, 59-99.

Kassin, S. M., Appleby, S. C., & Perillo, J. T. (2010). Interviewing suspects: Practice, science and future directions. *Legal and Criminological Psychology, 15*, 39-55.

Kassin, S. M., Drizin, S. A., Grisso, T., Gudjonsson, G. H., Leo, R. A., & Redlich, A. D. (2010). Police-induced confessions: Risk factors and recommendations. *Law and Human Behavior, 34*, 3-38.

Kassin, S. M., Goldstein, C. C., & Savitsky, K. (2003). Behavioral confirmation in the interrogation room: On the dangers of presuming guilt. *Law and Human Behavior, 27*, 187-203.

Köhnken, G. (1996). Social psychology and the law. In G. Semin & K. Fiedler (Eds.), *Applied social psychology* (pp. 257-282). London: Sage.

Köhnken, G., Milne, R., Memon, A., & Bull, R. (1999). The cognitive interview: A meta-analysis. *Psychology, Crime and Law, 5*, 3-27.

Köhnken, G., Schimossek, E., Aschermann, E., & Höfer, E. (1995). The cognitive interview and the assessment of the credibility of adults' statements. *Journal of Applied Psychology, 80*, 671-684.

Lamb, M. E., Sternberg, K. J., Esplin, P. W., Hershkowitz, I., Orbach, Y., & Hovav, M. (1997). Criterion-Based Content Analysis: a field validation study. *Child Abuse & Neglect, 21*, 255-264.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Leal, S., Vrij, A., Warmelink, L., Vernham, Z., & Fisher, R. P. (2015). You cannot hide your telephone lies: Providing a model statement as an aid to detect deception in insurance telephone calls. *Legal and Criminological Psychology, 20*, 129-146. doi:10.1111/lcrp.12017

Lee, K., Caverlee, J., & Webb, S. (2010). Uncovering social spammers: social honeypots+ machine learning. *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 435-442.

Levine, T. R., & McCornack, S. A. (2001). Behavioral adaptation, confidence, and heuristic-based explanations of the probing effect. *Human Communication Research, 27*, 471-502.

Levine, T. R., & McCornack, S. A. (2014). Theorizing about deception. *Journal of Language and Social Psychology, 33*, 431-440.

Levine, T. R., Serota, K. B., Carey, F., & Messer, D. (2013). Teenagers lie a lot: A further investigation into the prevalence of lying. *Communication Research Reports, 30*, 211-220. doi:10.1080/08824096.2013.806254

Lim, E. P., Nguyen, V. A., Jindal, N., Liu, B., & Lauw, H. W. (2010). Detecting product review spammers using rating behaviors. *Proceedings of the 19th ACM international conference on Information and knowledge management*, 939-948.

Lisak, D., Gardinier, L., Nicksa, S. C., & Cote, A. M. (2010). False allegations of sexual assault: An analysis of ten years of reported cases. *Violence Against Women, 16*, 1318-1334.

Maier, N. R. F., & Thurber, J. A. (1968). Accuracy of judgments of deception when an interview is watched, heard, and read. *Personnel Psychology, 21*, 23-30.

Mann, S., Vrij, A., & Bull, R. (2002). Suspect, lies and videotape: An analysis of authentic high-stakes liars. *Law and Human Behavior, 26*, 365-376.

Mann, S., Vrij, A., & Bull, R. (2004). Detecting true lies: Police officers' ability to detect suspects' lies. *Journal of Applied Psychology, 89*, 137-149.

Mann, S., Vrij, A., Fisher, R. P., & Robinson, M. (2008). See no lies, hear no lies: Differences in discrimination accuracy and response bias when watching or listening to police suspect interviews. *Applied Cognitive Psychology, 22*, 1062-1071.

Masip, J., Alonso, H., Garrido, E., & Antón, C. (2005). Generalized Communicative Suspicion (GCS) Among Police Officers: Accounting for the Investigator Bias Effect. *Journal of Applied Social Psychology, 35*, 1046-1066.

Masip, J., & Herrero, C. (2015). Police detection of deception: Beliefs about behavioral cues to deception are strong even though contextual evidence is more useful. *Journal of Communication, 65*, 125-145.

Masip, J., Sporer, A. L., Garido, E., & Herrero, C. (2005). The detection of deception with the reality monitoring approach: A review of the empirical evidence. *Psychology, Crime and Law, 11*, 99-122.

McDougall, A. J., & Bull, R. (2014). Detecting truth in suspect interviews: The effect of use of evidence (early and gradual) and time delay on Criteria-Based Content Analysis, Reality monitoring and inconsistency within suspect statements. *Psychology, Crime & Law*, 1-26.

Meijer, E. H., Verschuere, B., Gamer, M., Merckelbach, H., & Ben-Shakhar, G. (2016). Deception detection with behavioral, autonomic, and neural measures: Conceptual and methodological considerations that warrant modesty. *Psychophysiology*.

Memon, A., Meissner, C. A., & Fraser, J. (2010). The cognitive interview: A meta-analytic review and study space analysis of the past 25 years. *Psychology, Public Policy, and Law, 16*, 340-372.

Merckelbach, H. (2004). Telling a good story: fantasy proneness and the quality of fabricated memories. *Personality and Individual Differences, 37*, 1371-1382.

Merckelbach, H. (2015). Een "valse" aangifte. *De Psycholoog*, 40-48.

Morgan III, C. A., Colwell, K., & Hazlett, G. A. (2007). Efficacy of forensic statement analysis in distinguishing truthful from deceptive eyewitness accounts of highly stressful events. *Journal of forensic sciences, 56*, 1227-1234.

Nahari, G., & Vrij, A. (2014). Can I borrow your alibi? The applicability of the verifiability approach to the case of an alibi witness. *Journal of Applied Research in Memory and Cognition, 3*, 89-94.

Nahari, G., Vrij, A., & Fisher, R. P. (2012). Does the truth come out in the writing? SCAN as a lie detection tool. *Law and Human Behavior, 36*, 68-76. doi:10.1037/h0093965

Nahari, G., Vrij, A., & Fisher, R. P. (2014a). Exploiting liars' verbal strategies by examining the verifiability of details. *Legal and Criminological Psychology, 19*, 227-239. doi:10.1111/j.2044-8333.2012.02069.x

Nahari, G., Vrij, A., & Fisher, R. P. (2014b). The verifiability approach: Countermeasures facilitate its ability to discriminate between truths and lies. *Applied Cognitive Psychology, 28*, 122-128. doi:10.1002/acp.2974

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistics styles. *Personality and Social Psychology Bulletin, 29*, 665-675.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*, 175-220.

Nierop, N. M., van den Eshof, P., & Brandt, C. (2006). De beoordeling van geloofwaardigheid in zedenzaken: theorie en praktijk. *Nederlands Juristenblad, 43*, 2456-2464.

Niveau, G., Lacasa, M. J., Berclaz, M., & Germond, M. (2015). Inter-rater Reliability of Criteria-Based Content Analysis of Children's Statements of Abuse. *Journal of forensic sciences, 60*, 1247-1252.

Oberlader, V. A., Naefgen, C., Koppehele-Goseel, J., Quinten, L., Banse, R., & Schmidt, A. F. (2016). Validity of Content-Based Techniques to Distinguish True and Fabricated Statements: A Meta-Analysis. *Law and Human Behavior*, 1-63.

Ofshe, R. J., & Leo, R. A. (1997). The decision to confess falsely: Rational choice and irrational action. *Denver University Law Review, 74*, 979-1112.

Ott, M., Cardie, C., & Hancock, J. (2012). Estimating the prevalence of deception in online review communities. *Proceedings of the 21st international conference on World Wide Web*, 201-210.

Park, H. S., Levine, T., McCornack, S., Morrison, K., & Ferrara, M. (2002). How people really detect lies. *Communication Monographs, 69*, 144-157.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic Inquiry and Word Count: LIWC 2001*. Mahwah, NJ: Lawrence Erlbaum.

Porter, S., & ten Brinke, L. (2008). Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions. *Psychological Science, 19*, 508-514.

Porter, S., ten Brinke, L., & Wallace, B. (2012). Secrets and lies: Involuntary leakage in deceptive facial expressions as a function of emotional intensity. *Journal of Nonverbal Behavior, 36*, 23-37.

Porter, S., Woodworth, M., & Birt, A. R. (2000). Truth, lies, and videotape: An investigation of the ability of federal parole officers to detect deception. *Law and Human Behavior, 24*, 643-658. doi:10.1023/A:1005500219657

Porter, S., Yuille, J. C., & Lehman, D. R. (1999). The nature of real, implanted, and fabricated memories for emotional childhood events: Implications for the recovered memory debate. *Law and Human Behavior, 23*, 517-537.

Porter, S., & Yuille, Y. C. (1996). The language of deceit: An investigation of the verbal clues to deception in the interrogation context *Law and Human Behavior, 20*, 443-458.

Raskin, D. C., & Esplin, P. W. (1991). Statement validity assessment: Interview procedures and content analysis of children's statements of sexual abuse. *Behavioral Assessment, 13*, 265-291.

Rassin, E., Eerland, A., & Kuijpers, I. (2010). Let's find the evidence: An analogue study of confirmation bias in criminal investigations. *Journal of Investigative Psychology and Offender Profiling, 7*, 231-246.

Risinger, D. M., Saks, M. J., Thompson, W. C., & Rosenthal, R. (2002). The Daubert/Kumho implications of observer effects in forensic science: Hidden problems of expectation and suggestion. *California Law Review, 90*, 1-56.

Roma, P., San Martini, P., Sabatello, U., Tatarelli, R., & Ferracuti, S. (2011). Validity of Criteria-Based Content Analysis (CBCA) at trial in free-narrative interviews. *Child Abuse & Neglect, 35*, 613-620.

Ruby, C. L., & Brigham, J. C. (1997). The usefulness of the Criteria Based Content Analysis technique in distinguishing between truthful and fabricated allegations: A critical review. *Psychology, Public Policy, and Law, 3*, 705-737.

Santtila, P., Roppola, H., Runtti, M., & Niemi, P. (2000). Assessment of child witness statements using Criteria-Based Content Analysis (CBCA): The effects of age, verbal ability, and interviewer's emotional style. *Psychology, Crime and Law, 6*, 159-179

Sapir, A. (2005). *The LSI course on scientific content analysis (SCAN)*. Phoenix, AZ: Laboratory for Scientific Interrogation.

Schelleman-Offermans, K., & Merckelbach, H. (2010). Fantasy proneness as a confounder of verbal lie detection tools. *Journal of Investigative Psychology and Offender Profiling, 7*, 247-260.

Schooler, J. W., Gerhard, D., & Loftus, E. F. (1986). Qualities of the unreal. *journal of experimental Psychology, 12*, 171-181.

Serota, K. B., Levine, T. R., & Boster, F. J. (2010). The prevalence of lying in America: Three studies of self-reported lies. *Human Communication Research*, 2-25. doi:10.1111/j.1468-2958.2009.01366.x

Shafir, E. (1995). Compatibility in cognition and decision. *Psychology of Learning and Motivation, 32*, 247-274.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428.

Smith, N. (2001). Reading between the lines: an evaluation of the Scientific content Analysis technique (SCAN). *Police Research Series Paper 135*, 1-42.

Snook, B., Cullen, R. M., Bennell, C., Taylor, P. J., & Gendreau, P. (2008). The criminal profiling illusion what's behind the smoke and mirrors? *Criminal Justice and Behavior, 35*, 1257-1276.

Sooniste, T., Granhag, P. A., Knieps, M., & Vrij, A. (2013). True and false intentions: asking about the past to detect lies about the future. *Psychology, Crime and Law, 19*, 673-865. doi:10.1080/1068316X.2013. 793333

Sporer, S. L. (1997). The less travelled road to truth: Verbal cues in deception detection in accounts of fabricated and self-experienced events. *Applied Cognitive Psychology, 11*, 373-397.

Sporer, S. L. (2004). Reality monitoring and detection of deception. In P. A. Granhag & L. A. Strömwall (Eds.), *The detection of deception in forensic contexts* (pp. 64-102). UK: Cambridge: University press.

Sporer, S. L., Masip, J., & Cramer, M. (2014). Guidance to detect deception with the Aberdeen Report Judgment Scales: are verbal content cues useful to detect false accusations? *The American Journal of Psychology, 127*, 41-63.

Sporer, S. L., & Schwandt, B. (2007). Moderators of nonverbal indicators of deception. *Psychology, Public Policy, and Law, 13*, 1-34.

Steller, M. (1989). Recent developments in statement analysis. In J. C. Yuille (Ed.), *Credibility Assessment* (pp. 135-154). Dordrecht: Kluwer Academic Publishers.

Steller, M., & Köhnken, G. (1989). Criteria Based Statement Analysis. In D. C. Raskin (Ed.), *Psychological methods in criminal investigation and evidence* (pp. 217-245). New York: Springer

Stringhini, G., Kruegel, C., & Vigna, G. (2010). Detecting spammers on social networks. *Proceedings of the 26th Annual Computer Security Applications Conference*, 1-9.

Strömwall, L. A., & Granhag, P. A. (2003). How to detect deception? Arresting the beliefs of police officers, prosecutors and judges. *Psychology, Crime and Law, 9*, 19-36.

Strömwall, L. A., Granhag, P. A., & Hartwig, M. (2004). Practicioners' beliefs about deception. In L. A. Strömwall & P. A. Granhag (Eds.), *The detection of deception in forensic contexts*. Cambridge: Cambridge University Press.

Suckle-Nelson, J. A., Colwell, K., Hiscock-Anisman, C. K., Florence, S., Youschak, K. e., & Duarte, A. (2010). Assessment Criteria Indicative of Deception (ACID): Replication and gender differences. *The Open Criminology Journal, 3*, 23-30.

Swets, J. A. (2014). *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers.*: Psychology Press.

Taylor, R., & Hick, R. F. (2007). Believed cues to deception: Judgments in self-generated trivial and serious situations. *Legal and Criminological Psychology, 12*, 321-331.

The Global Deception Research Team. (2006). A world of lies. *Journal of Cross-Cultural Psychology, 37*, 60-74.

Trankell, A. (1972). *Reliability of evidence*. Stockholm, Sweden: Beckmans.

Undeutsch, U. (1967). Beurteilung der glaubhaftigkeit von Aussagen. In U. Undeutsch (Ed.), *Handbuch der psychologie Vol 11: Forensische Psychologie*. Göttingen, Germany: Hogrefe.

van Amelsvoort, A., Rispens, I., & Grolman, H. (2015). *Handleiding verhoor [Manual interrogations]*. Amsterdam: Reed Business.

Van Geest, E. (2008). *Misleidingen in verklaringen: De SCAN-techniek*. Katholieke Universtiteit Leuven.

van Koppen, P. J. (2010). De psycholoog en het deskundigenbewijs. In P. J. van Koppen, H. Merckelbach, M. Jelicic, & J. W. de Keijser (Eds.), *Reizen met mijn rechter: Psychologie van het recht* (pp. 401-422). Deventer: Kluwer.

Vanderhallen, M., Jaspaert, E., & Vervaeke, G. (2015). SCAN as an investigative tool. *Police Practice and Research*, 1-15.

Vrij, A. (2005). Criteria Based Content Analysis: A qualitative review of the first 37 studies. *Psychology, Public Policy, and Law, 11*, 3-41.

Vrij, A. (2008a). Beliefs about nonverbal and verbal cues to deception. In A. Vrij (Ed.), *Detecting lies and deceit* (pp. 115-140). Chirchester: Wiley.

Vrij, A. (2008b). *Detecting lies and deceit: Pitfalls and opportunities*. Chichester: Wiley.

Vrij, A. (2008c). Nonverbal dominance versus verbal accuracy in lie detection: A plea to change police practice. *Criminal Justice and Behavior, 35*, 1323-1336.

Vrij, A. (2008d). Reality Monitoring. In G. Davies & R. Bull (Eds.), *Detecting lies and deceit: Pitfalls and opportunities* (pp. 261-280). Chichester: Wiley.

Vrij, A., Akehurst, L., & Knight, S. (2006). Police officers', social workers', teachers' and the general public's beliefs about deception in children, adolescents and adults. *Legal and Criminological Psychology, 11*, 297-312.

Vrij, A., Akehurst, L., Soukara, S., & Bull, R. (2002). Will the truth come out? The effect of deception, age, status, coaching, and social skills on CBCA scores. *Law and Human Behavior, 26*, 261-283.

Vrij, A., Akehurst, L., Soukara, S., & Bull, R. (2004a). Detecting deceit via analysis of verbal and nonverbal behavior in children and adults. *Human Communication Research, 30*, 8-41.

Vrij, A., Akehurst, L., Soukara, S., & Bull, R. (2004b). Let me inform you how to tell a convincing story: CBCA and Reality Monitoring scores as a function of age, coaching and deception. *Canadian Journal of Behavioural Science, 36*, 113-126.

Vrij, A., Edward, K., Roberts, K. P., & Bull, R. (2000). Detecting deceit via analysis of verbal and nonverbal behavior. *Journal of Nonverbal Behavior, 24*, 239-263.

Vrij, A., Fisher, R., Mann, S., & Leal, S. (2006). Detecting deception by manipulating cognitive load. *Trends in Cognitive Sciences, 10*, 141-142.

Vrij, A., Fisher, R., Mann, S., & Leal, S. (2008). A cognitive load approach to lie detection. *Journal of Investigative Psychology and Offender Profiling, 5*, 39-43.

Vrij, A., & Granhag, P. A. (2012). Eliciting cues to deception and truth: What matters are the questions asked. *Journal of Applied Research in Memory and Cognition, 1*, 110-117. doi:10.1016/j.jarmac.2012.02.004

Vrij, A., Granhag, P. A., Mann, S., & Leal, S. (2011). Outsmarting the liars: Toward a cognitive lie detection approach. *Current Directions in Psychological Science, 20*, 28-32.

Vrij, A., Granhag, P. A., & Porter, S. (2010). Pitfalls and opportunities in nonverbal and verbal lie detection. *Psychological Science in the Public Interest, 11*, 89-121.

Vrij, A., Kneller, W., & Mann, S. (2000). The effect of informing liars about Criteria Based Content Analysis on their ability to deceive CBCA raters. *Legal and Criminological Psychology, 5*, 57-70.

Vrij, A., Leal, S., Granhag, P. A., Mann, S., Fisher, R., Hillman, J., & Sperry, K. (2009). Outsmarting the liars: The benefit of asking unanticipated questions. *Law and Human Behavior, 33*, 159-166. doi:10.1007/s10979-008-9143-y

Vrij, A., Leal, S., Mann, S., & Fisher, R. (2012). Imposing cognitive load to elicit cues to deceit: inducing the reverse order technique naturally. *Psychology, Crime and Law, 18*, 579-594.

Vrij, A., & Mann, S. (2001). Who killed my relative? Police officers' ability to detect real-life high stake lies. *Psychology, Crime and Law, 7*, 119-132.

Vrij, A., & Mann, S. (2006). Criteria Based Content Analysis: An empirical test of its underlying processes. *Psychology, Crime and Law, 12*, 337-349.

Vrij, A., Mann, S., & Fisher, R. (2006). Information-gathering vs accusatory interview style: Individual differences in respondents' experiences. *Personality and Individual Differences, 41*, 589-599.

Vrij, A., Mann, S., Fisher, R., Leal, S., Milne, R., & Bull, R. (2008). Increasing cognitive load to facilitate lie detection: The benefit of recalling an event in reverse order. *Law and Human Behavior, 32*, 253-265.

Vrij, A., Mann, S., & Leal, S. (2008). Reality Monitoring. In L. Smets & A. Vrij (Eds.), *Het analyseren van de geloofwaardigheid van verhoren: Het gebruik van leugendetectiemethoden* (pp. 81-84). Brussel: Politeia.

Vrij, A., Nahari, G., Isitt, R., & Leal, S. (2016). Using the verifiability lie detection approach in an insurance claim setting. *Journal of Investigative Psychology and Offender Profiling*.

Vrij, A., & Semin, G. (1996). Lie experts' beliefs about nonverbal indicators of deception. *Journal of Nonverbal Behavior, 20*, 65-80.

Warmelink, L., Vrij, A., Mann, S., Jundi, S., & Granhag, P. A. (2012). The effect of question expectedness and experience on lying about intentions. *Acta Psychologica, 141*, 178-183.

Wegener, H. (1989). The present state of statement analysis. In J. C. Yuille (Ed.), *Credibility Assessment* (pp. 121-134). Dordrecht: Kluwer Academic Publishers.

Wu, G., Greene, D., Smyth, B., & Cunningham, P. (2010). Distortion as a validation criterion in the identification of suspicious reviews. *Proceedings of the First Workshop on Social Media Analytics*, 10-13.

Zijlstra, H. M., Middendorp, H., Meerveld, T., & Geenen, R. (2005). Validiteit van de Nederlandse versie van de [Validity of the Dutch version of] Linguistic Inquiry and Word Count (LIWC). *Nederlands Tijdschrift voor de Psychologie [Dutch Journal of Psychology), 60*, 55-63.


Maastricht Police Department
January 27, 2017
Statement of John Smith
I live in a two-floor house in Canterbury. I rent the house with a few other boys, because it is very expensive in this neighbourhood. However, last night all the boys were out except for me. Somewhere after midnight I was woken up by a noise. The noise wasn't very loud but apparently it was loud enough to wake me up. I tried to listen more carefully, but I had no idea what it was. So I got out of bed and looked out of the window and at first I couldn't work out what was causing the noise. My bedroom is upstairs at the front of the house and it seemed to be coming from directly below me. I leaned out of the window so I could see better and I saw two men dressed in black doing something to the downstairs window. As I leaned out they climbed through the window and into the front room. I didn't get a very good look at them here but they were both of average build and looked in pretty good shape. They both wore black hats. I looked around to see if there were any more but I did not see anything noteworthy. After a little while I noticed there

Summaries

ENGLISH

This dissertation deals with the evaluation of the three verbal credibility assessment methods, namely Scientific Content Analysis (SCAN), Criteria Based Content Analysis (CBCA) and Reality Monitoring (RM). Part 1 primarily focused on investigating SCAN and its properties. Chapter 2 examined beliefs about verbal and nonverbal cues to deception and how these beliefs could be related to the popularity of SCAN. It was shown that both police officers and laypersons hold incorrect beliefs about the diagnosticity of nonverbal cues, but were less inclined to overestimate the relationship between verbal cues and deception. Contrary to what we expected, results indicated no strong endorsement of SCAN items.

Chapter 3 investigated the usage of SCAN in practice by examining 82 sexual abuse cases of the Dutch police in which SCAN had been applied. We carefully examined which SCAN criteria were identified by the Amsterdam police SCAN analysts in the written statements of these cases. The results of this study showed that SCAN is largely driven by a set of 12 criteria, and these criteria served as a basis for our subsequent studies. However, although the SCAN analysts who coded the SCAN criteria in these cases were trained in using SCAN,

Chapter 4 examined whether SCAN was able to accurately discriminate between true and fabricated statements about a recent negative autobiographical event. All statements were analysed using the criteria derived from Chapter 3. Results showed no empirical support for SCAN as a lie detection tool. Possible reasons for the lack of SCAN's discriminability, such as the absence of criteria checking various types of details, are discussed.

The remainder of this dissertation deals with the merits of SCAN in comparison with those of CBCA and RM, and investigates boundary conditions and possible improvements for all three methods. Chapter 5 addressed the important issue of contextual information and how these tools might be influenced by credibility enhancing or reducing information. We found no difference between the control group and the groups that relied on the CBCA, RM or SCAN to evaluate statements. More precisely, statements preceded by positive context information were found to be richer in criteria than statements preceded by negative information, regardless of whether tools were used or not. These findings suggest that these methods do little to decrease the influence of biasing context information, thereby highlighting the importance of blind raters when assessing the credibility of statements.

On the other hand, in light of the methods' sensitivities to manipulations, in Chapter 6 we tested whether giving individuals an example of a detailed statement, fulfilling all relevant criteria of the CBCA, RM, and SCAN, without explaining these criteria would influence the methods' scores. We expected that this effect would be more pronounced in truth tellers than in liars, thereby increasing discriminability of these methods. First, results clearly indicated that these lie detection methods differ in their sensi-

tivity to measuring truthfulness. Both RM and CBCA could accurately discriminate between true and fabricated statements, while SCAN was not. Second, contrary to our expectations, supplying individuals with a detailed example statement did not increase or decrease discriminability of the verbal credibility assessment.

In the final Chapter, theoretical and practical implications of the reported findings are discussed, with an emphasis on current developments in lie detection research. To conclude, this dissertation has established that lies differ observably from truthful statements, both in their content and linguistic features and that it is possible to highlight these differences with the aid of scientifically based verbal credibility assessment methods. SCAN, however, does not fall into this category.

NEDERLANDS

Dit proefschrift evalueert drie verbale analyse-methoden, namelijk de Scientific Content Analysis (SCAN), Criteria Based Content Analysis (CBCA) en Reality Monitoring (RM). Deel 1 behandelt voornamelijk onderzoek naar SCAN en haar psychometrische eigenschappen. In Hoofdstuk 2 onderzochten we (1) welke opvattingen politiebeambten en leken hebben over verbale en non-verbale leugendetectiesignalen en (2) of deze opvattingen de populariteit van SCAN kunnen verklaren. Zoals verwacht hadden zowel politiebeambten als leken veelal foutieve opvattingen over welke non-verbale signalen bruikbaar zijn om leugens te detecteren. Beide groepen bleken echter wel meer juiste opvattingen te hebben over verbale signalen. Wanneer we vroegen naar hun opvattingen over de SCAN-hypothesen, bleken beide groepen deze hypothesen doorgaans niet te volgen. We vonden hiermee geen bewijs dat de populariteit van SCAN uitgelegd kan worden door haar attractieve hypothesen.

In Hoofdstuk 3 bestudeerden we 82 politiedossiers waarin de SCAN werd aangewend. In deze dossiers scoorden wij de aan- of afwezigheid van de verscheidene SCAN-criteria om inzicht te krijgen in hoe de SCAN door politiebeambten wordt toegepast. Hieruit concludeerden wij dat SCAN–oordelen grotendeels berusten op een set van 12 criteria. Deze set diende als basis voor onze verdere studies naar de validiteit van de SCAN. We onderzochten ook of de SCAN-criteria wel betrouwbaar gecodeerd konden worden. Dit bleek niet het geval. De betrouwbaarheid tussen de beoordelaars was behoorlijk laag, met een gemiddelde overeenkomst van slechts 31%. Deze lage overeenkomst is opvallend, aangezien alle beoordelaars de driedaagse SCAN-training succesvol hadden afgerond.

Hoofdstuk 4 beantwoordt de vraag of de SCAN in staat is om onderscheid te maken tussen echte en verzonnen verklaringen. 117 ware en 117 onware verklaringen werden geanalyseerd met de criteria ontleend aan Hoofdstuk 3. SCAN bleek – met een gemiddelde accuraatheid van 52% - geen onderscheid te kunnen maken tussen ware en onware verklaringen. Verscheidene redenen kunnen dit onvermogen uitleggen. Zo verschillen de interpretaties van sommige SCAN-criteria van die van CBCA en RM (i.e., leugenachtigheid voor SCAN en waarheid voor CBCA en RM) en houdt de SCAN geen rekening met verschillende soorten kenmerkende details (e.g., perceptuele details), terwijl deze volgens de literatuur juist belangrijk zijn bij het herkennen van leugenaars.

Het vervolg van dit proefschrift gaat over de merites van de SCAN in vergelijking met CBCA en RM, en onderzoekt randvoorwaarden en mogelijke verbeteringen voor deze methoden. Hoofdstuk 5 richt zich op de belangrijke kwestie van contextuele informatie en hoe verbale analyse methoden hierdoor kunnen worden beïnvloed. Verklaringen die werden voorafgegaan door contextuele informatie om de geloofwaardigheid van de aangeefster te verhogen (e.g., getuigen beschrijven verdachte als opvliegend en agressief) bleken rijker te worden gescoord aan criteria dan verklaringen voorafgegaan door informatie om de geloofwaardigheid van de aangeefster te verminderen (e.g., aangeef-

ster heeft problemen met moeder vanwege haar rebels gedrag). Dit gold in dezelfde mate voor alle drie de methoden, hetgeen suggereert dat deze methoden weinig of niets doen om de invloed van contextuele informatie te temperen. Deze bevindingen benadrukken het belang van blinde beoordelaars bij het onderzoeken van de geloofwaardigheid van verklaringen.

In hoofdstuk 6 testen we of de accuraatheid van CBCA, RM en SCAN verbeterd kan worden door het gebruik van een gedetailleerde voorbeeldverklaring. De helft van de proefpersonen kreeg vooraf aan het geven van hun verklaring een voorbeeldverklaring te lezen die aan alle relevante criteria van de CBCA, RM en SCAN voldeed, de andere helft kreeg dit voorbeeld niet. Zoals verwacht waren de RM en CBCA in staat om ware en onware verklaringen van elkaar te onderscheiden, terwijl dit voor SCAN niet het geval was. In tegenstelling tot onze verwachtingen zorgde het gebruik van een voorbeeldverklaring niet voor een betere discriminatie tussen ware en onware verklaringen maar resulteerde het wel in langere verklaringen. Dit betekent dat men meer geneigd is om informatie te geven na het lezen van een gedetailleerde verklaring. Het verkrijgen van informatie is een belangrijk doel in opsporingsonderzoek en de eenvoudige manipulatie van het geven van een voorbeeldverklaring lijkt hierbij te helpen.

In het laatste hoofdstuk wordt de theoretische en praktische betekenis van de gerapporteerde bevindingen besproken, waarbij de nadruk wordt gelegd op recente ontwikkelingen in leugendetectie-onderzoek. Dit proefschrift, tot slot, ondersteunt voorgaand onderzoek in dat leugens waarneembaar verschillen van waarheidsgetrouwe verklaringen. Zowel in inhoudelijke kenmerken als in taalkundige eigenschappen en dat het mogelijk is om deze verschillen aan de hand van wetenschappelijk onderbouwde verbale analysemethoden te toetsen. SCAN behoort echter niet tot deze methoden.

Valorisation Addendum

In the following valorisation addendum, five issues will be addressed regarding the current dissertation "*Catching liars by listening carefully: Promises and challenges for credibility assessment*". The first point will address the *relevance* of the present dissertation. The second issue is concerned with the *target group* to whom the results of this dissertation may be relevant. The third issue deals with *specific activities and services* that may be derived based on the results presented in this dissertation. The fourth issue explains the *innovative approach* and results. The last issue explains how the results of this dissertation can be *implemented in practices* and which opportunities there are. I want to note that the relevance and innovative nature of the findings presented in this thesis have already been highlighted in various chapters. Nonetheless, I will shortly describe them below.

RELEVANCE OF THE RESEARCH

Investigative authorities are often confronted with deceptive suspects, with witnesses and bogus victims who raise dubious claims (e.g., fabricate traumatic stories, false insurance claims) (Greer, 2000; Gudjonsson, Sigurdsson, Asgeirsdottir, & Sigfusdottir, 2007; Lisak, Gardinier, Nicksa, & Cote, 2010). Furthermore, the first DNA exonerations in the 1990's have shown that around 20-25% of the wrongful imprisonments resulted from false confessions (Kassin, 2012), and thus the inability of detectives to recognize these types of lies. For example, in 2103, detective Hero Brinkman interrogated a woman who was suspected of hitting another woman in the face with a broken glass. The suspect confessed to this crime, however, she had nothing to do with the incident. Brinkman did not recognize her confession as a false one. Even more so, a few months later, the actual perpetrator stepped forward and confessed to the crime. Nonetheless, Brinkman suspected her of giving a false confession and sent her home (Haenen, 2016). This is only one example in which police detectives failed to accurately detect truths and lies.

Indeed, research has shown the detection of deception to be challenging. In general, people, including trained police officers, only perform around chance level when detecting deception (Aamodt & Custer, 2006; Bond & DePaulo, 2006, 2008; Vrij & Mann, 2001). One reason is that people rely heavily on nonverbal cues when making deception verdicts (for an overview see Vrij, 2008a; chapter 2 of this dissertation), regardless of a large body of research showing that deception cannot be reliably inferred from behavior (DePaulo et al., 2003; Sporer & Schwandt, 2007). Instead, it is better to rely on the verbal content when making credibility judgments (Hauch, Sporer, Michael, & Meissner, 2014; Masip, Alonso, Garrido, & Antón, 2005; Vrij, 2005, 2008b). Moreover, meta-analytic research reported a higher lie detection accuracy if the training was based on verbal cues compared with nonverbal training (Hauch et al., 2014). Consequently, con-

tent should accordingly be favored over behavior (Levine & McCornack, 2014; Masip & Herrero, 2015; Vrij, 2008b).

Several veracity assessment methods have been developed that rely specifically on the content of a statement, such as Criteria-Based Content Analysis (CBCA; Steller & Köhnken, 1989), Reality Monitoring (RM; Johnson & Raye, 1981) and the Scientific Content Analysis (SCAN; Sapir, 2005). This dissertation examined the effectiveness of these verbal credibility assessment methods; more precisely it evaluated the usefulness of SCAN as a lie detection method and investigated boundary conditions and possible improvements for verbal credibility assessment.

TARGET GROUPS

The findings of the current dissertation are relevant for law enforcement agencies, legal professionals and policy makers. The present work clearly shows that SCAN should not be applied by investigate authorities as a tool to assess witness', victims' or suspects' credibility. First, SCAN's inter-rater reliability was found to be disappointingly low, with an average agreement of 31%. Second, SCAN criteria were unable to accurately discriminate between truthful and fabricated statements, showing SCAN is insufficiently developed as an investigate tool. Furthermore, current findings have highlighted that even though RM and CBCA can be reliably used as a lie detection tool, and their accuracy is significantly above chance level, these methods do little to decrease the influence of biasing context information. This stresses the importance of blind raters when assessing the credibility of statements.

Furthermore, these results are relevant for the academic community and especially researchers in the field of psychology and law, deception detection and communication. We hope that the scientific community will replicate and extend our findings presented in the current dissertation.

ACTIVITIES AND SERVICES

This thesis has established that it is possible to distinguish between truthful and fabricated statements based on their content and linguistic features, and that it is possible to highlight these differences with the aid of verbal credibility assessment methods, such as CBCA and RM. However, the findings from this thesis emphasize that SCAN does not fulfil the necessary psychometric standards to be applied as a lie detection tool. SCAN has a weak criterion and construct validity and its inter-rater reliability is disappointingly low. Therefore, investigative authorities have been informed about these findings and have been advised to refrain from using SCAN in the future.

Additionally, the findings presented in the various chapters in this dissertation indicate that people - including police officers - still hold wrongful beliefs about which cues are diagnostic for detection deception, that verbal credibility assessment is sensitive to confirmation bias, and that a more active approach of obtaining statements shows potential for credibility assessment. Consequently, the insights gained by this line of research might be used to improve the diagnostic accuracy of verbal credibility assessment tools and to develop novel deception detection trainings primarily aimed at informing investigative authorities to shift their attention from non-verbal to more diagnostic verbal cues.

INNOVATION OF THE RESEARCH

The work presented in the current dissertation is one among the first peer-reviewed studies that investigated the validity of SCAN. Given the unstandardized nature of the SCAN method, the present dissertation first examined which SCAN criteria are most frequently used in the field. This resulted in a list of 12 unique criteria that were used a basis for further investigation of SCAN. Next, the validity of SCAN as a whole, as well as the validity of separate SCAN criteria, were investigated in multiple studies (chapters 4, 5 and 6).

Furthermore, although RM and CBCA have been widely researched in the last three decades, the current dissertation extended previous literature by investigating potentially boundary conditions and improvements for these methods. For example, information about how vulnerable these methods are to contextual bias is relevant in the light of guidelines concerning the handling of this extra-domain information. Therefore, Chapter 5 investigated to what extent credibility assessment methods were influenced by contextual information. Our findings demonstrated that contextual information influenced subsequent credibility assessment.

This dissertation also focused on improving CBCA, RM and SCAN's accuracy by supplying participants with a model statement. Our results showed that the model statement did not increase CBCA, RM and SCAN's accuracy, but led to other cues, which can serve a valuable addition for lie detection research.

KNOWLEDGE DISSEMINATION

The outcomes of the studies described in this dissertation have all been communicated in several ways. The studies presented in chapters 2 to 6 are all published in international peer reviewed journals. Furthermore, these chapters have been presented at several national and international conferences, which were attended by both researchers and legal professionals from around the world. Chapters 1 and 7 are adapted and

translated versions of manuscripts that were published in Dutch and Belgian journals. Also, several of the studies have evoked interest outside the scientific community, and have been presented at the annual “recherchekundige bijeenkomst Limburg”. Additionally, knowledge valorisation is and will continue to be stimulated by communicating findings through journal publications and conference presentation and other invited presentations (e.g., information sessions for judges, police detectives, police academy).

REFERENCES

Aamodt, M. G., & Custer, H. (2006). Who can best catch a liar? A meta-analysis of individual differences in detecting deception. *Forensic Examiner, 15*, 6-11.

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Individual Differences, 10*, 214-234.

Bond, C. F., & DePaulo, B. M. (2008). Individual differences in judging deception: Accuracy and bias. *Psychological Bulletin, 134*, 477-492.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin, 129*, 74-118.

Greer, E. (2000). The truth behind legal dominance feminism's "two percent false rape claim" figure. *Loyola Of Los Angeles Law Review, 33*, 947-972.

Gudjonsson, G. H., Sigurdsson, J. F., Asgeirsdottir, B. B., & Sigfusdottir, I. D. (2007). Custodial interrogation: What are the background factors associated with claims of false confession to police? *Journal of Forensic Psychiatry & Psychology, 18*, 266-275.

Haenen, M. (2016). Oud-PVV'er Hero Brinkman bij politie gedegradeerd. Retrieved from http://www.nrc.nl/.

Hauch, V., Sporer, S. L., Michael, S. W., & Meissner, C. A. (2014). Does training improve the detection of deception? A meta-analysis. *Communication Research*, 1-61.

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review, 88*, 67-85.

Kassin, S. M. (2012). Why confessions trump innocence. *American Psychologist, 67*, 431.

Levine, T. R., & McCornack, S. A. (2014). Theorizing about deception. *Journal of Language and Social Psychology, 33*, 431-440.

Lisak, D., Gardinier, L., Nicksa, S. C., & Cote, A. M. (2010). False allegations of sexual assault: An analysis of ten years of reported cases. *Violence Against Women, 16*, 1318-1334.

Masip, J., Alonso, H., Garrido, E., & Antón, C. (2005). Generalized Communicative Suspicion (GCS) Among Police Officers: Accounting for the Investigator Bias Effect. *Journal of Applied Social Psychology, 35*, 1046-1066.

Masip, J., & Herrero, C. (2015). Police detection of deception: Beliefs about behavioral cues to deception are strong even though contextual evidence is more useful. *Journal of Communication, 65*, 125-145.

Sapir, A. (2005). *The LSI course on scientific content analysis (SCAN)*. Phoenix, AZ: Laboratory for Scientific Interrogation.

Sporer, S. L., & Schwandt, B. (2007). Moderators of nonverbal indicators of deception. *Psychology, Public Policy, and Law, 13*, 1-34.

Steller, M., & Köhnken, G. (1989). Criteria Based Statement Analysis. In D. C. Raskin (Ed.), *Psychological methods in criminal investigation and evidence* (pp. 217-245). New York: Springer

Vrij, A. (2005). Criteria Based Content Analysis: A qualitative review of the first 37 studies. *Psychology, Public Policy, and Law, 11*, 3-41.

Vrij, A. (2008a). Beliefs about nonverbal and verbal cues to deception. In A. Vrij (Ed.), *Detecting lies and deceit* (pp. 115-140). Chirchester: Wiley.

Vrij, A. (2008b). Nonverbal dominance versus verbal accuracy in lie detection: A plea to change police practice. *Criminal Justice and Behavior, 35*, 1323-1336.

Vrij, A., & Mann, S. (2001). Who killed my relative? Police officers' ability to detect real-life high stake lies. *Psychology, Crime and Law, 7*, 119-132.


Maastricht Police Department
January 27, 2017
Statement of John Smith
I live in a two-floor house in Canterbury. I rent the house with a few other boys, because it is very expensive in this neighbourhood. However, last night all the boys were out except for me. Somewhere after midnight I was woken up by a noise. The noise wasn't very loud but apparently it was loud enough to wake me up. I tried to listen more carefully, but I had no idea what it was. So I got out of bed and looked out of the window and at first I couldn't work out what was causing the noise. My bedroom is upstairs at the front of the house and it seemed to be coming from directly below me. I leaned out of the window so I could see better and I saw two men dressed in black doing something to the downstairs window. As I leaned out they climbed through the window and into the front room. I didn't get a very good look at them here but they were both of average build and looked in pretty good shape. They both wore black hats. I looked around to see if there were any more but I did not see anything noteworthy. After a little while I noticed there

Dankwoord

Het moment is eindelijk aangebroken dat ik kan beginnen te schrijven aan mijn dankwoord. De woorden die ik hier wil neerschijven zijn al vaker door mijn hoofd gegaan terwijl ik kritische reviews te verwerken kreeg, of wanneer er weer onverwachte problemen de kop op staken tijdens mijn onderzoek. Op zulke momenten verlang je natuurlijk naar het bereiken van maar één doel, het proefschrift! Maar, een proefschrift, dat schrijf je vanzelfsprekend niet alleen. De afgelopen jaren hebben heel wat mensen mij geholpen en daarom wil ik hen graag bedanken voor hun steun, hun kritiek, hun bemoedigende woorden en alle gezelligheid die dat met zich mee bracht.

In de eerste plaats ben veel dank verschuldigd aan jou, Ewout. In de afgelopen jaren heb ik ongelofelijk veel van jou geleerd. Je hebt me tijdens mijn studie al laten zien hoe leuk onderzoek eigenlijk kan zijn. Altijd kon (en kan) ik bij jou binnen lopen om te overleggen over vreemde resultaten, kritische reviews, creatieve oplossingen en andere inhoudelijke en schrijf technische tips. Kortom, eigenlijk heb ik je deur gewoon platgelopen. Jouw enthousiasme voor het onderzoek en kennis van de literatuur zorgden ervoor dat ik steeds met goede moed aan de slag kon gaan mijn onderzoek. Weet dat ik je harde werk, je toewijding en steun in de afgelopen jaren erg heb gewaardeerd. Ik wil je daarom ook enorm bedanken voor het vertrouwen dat je mij de afgelopen jaren hebt gegeven en ik hoop dat we onze samenwerking nog lang kunnen voortzetten.

Dankjewel, Aldert, voor je betrokkenheid bij mijn project. Jouw deskundigheid en enthousiasme hebben mij enorm geholpen bij het schrijven van dit proefschrift. Ook jouw snelle antwoord op de manuscripten die ik je toestuurde heb ik erg gewaardeerd, evenals natuurlijk het inhoudelijke commentaar.

Harald, bedankt voor al je gedetailleerde commentaar en objectieve blik op mijn manuscripten. Ondanks je drukke agenda was geen enkele vraag je te veel en dat waardeer ik enorm. Ik bewonder je kennis, je enthousiasme voor wetenschappelijk onderzoek, en bovenal, je aangename, warme persoonlijkheid.

Mijn twee paranimfen, Linsey en Jill, jullie betekenen zoveel meer voor mij dan gewoon (ex)collega's. Ik keek/kijk altijd erg uit naar onze lunches en gezellige uitjes, ook buiten het werk om. We hebben samen zoveel leuke (en minder leuke) momenten meegemaakt en altijd kan ik bij jullie terecht. Ik ben daarom ook enorm blij dat jullie op mijn promotie dag aan mijn zijde staan. Ik hoop dat onze paden zich in de toekomst nog vaak mogen blijven kruisen. Dank jullie wel, lieve vriendinnen.

Lieve collega's, ook jullie bedankt voor al die gezellige lunches, leuke feestjes en natuurlijk de aangename sfeer op de werkvloer. Ex-kamergenootjes Dalena, Elly en Colinda, ook aan jullie bedankt, want zonder onze gezellige kletsmomentjes was promoveren lang niet zo leuk geweest. Anna, you rock! You are such a kind and loving person, and I

will miss you a lot when we are no longer roommates. Thank you for all the times you helped me out, and just for listening to all my crazy stories. Pim, ongelofelijk bedankt voor al die keren dat je geduldig naar mijn frustraties hebt moeten luisteren. Thomas, bedankt voor al je hulp, of het nu over statistiek ging, vreemde resultaten of vage onderzoeks-ideeën, jij wist altijd raad. Dank aan Tom voor de spannende pingpong wedstrijden, en dan natuurlijk vooral de sets die ik mocht winnen! Thank you Alana, Alfons, Anna, Brianna, Conny, Colinda, Corine, David, Elly, Ewout, Harald, Henry, Irena, Ivan, Jill, Katherine, Kim, Linsey, Maarten, Maartje, Marko, Melanie, Nathalie, Nina, Robin, Sergii, Tameka, Thomas and Tom!

Veel dank gaat ook uit naar de studenten en student assistenten die me in de afgelopen jaren hebben geholpen met het verzamelen van gegevens. Linda Schoppink, Vivien Borbély, Sanne Hellemons, Esther Brink, Marilien Marzolla en Yves Moerskofski, bedankt voor jullie enthousiasme en inzet! Het betekende veel voor me.

Tot slot, dank aan mijn lieve familie en vrienden voor al jullie steun. Pap, Nick, mam en Jacky, bedankt om er altijd voor me te zijn. Lieve pap en mam, zonder jullie zou ik deze kans nooit hebben gehad. Bedankt voor jullie onvoorwaardelijke liefde, steun en onuitputtelijk geduld. Nick, bedankt om mijn beste vriend te zijn. Je staat altijd voor me klaar, zeker wanneer ik het moeilijk heb en dat waardeer ik enorm. Bedankt om alles zo goed te kunnen relativeren en om me steeds met je gevatte opmerkingen aan het lachen te brengen. Sofie, ook een dikke merci om mijn beste vriendin te willen zijn. Met jou erbij is alles gewoon leuker!

Curriculum Vitae

Glynis Bogaard werd geboren op 14 juni 1987 te Bree. In juni 2005 behaalde zij haar ASO diploma aan het Instituut Heilig Graf te Kinrooi. In september van dat jaar begon zij aan haar studie Psychologie aan de Universiteit Maastricht en in augustus 2009 behaalde ze haar masterdiploma in Psychology and Law *cum laude*. Vanaf 1 november 2009 werd zij aan de Faculteit der Psychologie en Neurowetenschappen van de Universiteit Maastricht aangesteld als promovenda voor onderzoek naar verbale leugendetectiemethoden.

Glynis Bogaard was born on June 14, 1987 in Bree. In June 2005, she finished her secondary school at the Insituut Heilig Graf in Kinrooi. In September of that year, she started studying Psychology at Maastricht University and in August 2009 she received her masters' diploma in Psychology and Law, *cum laude*. From November 2009, she was appointed at the Faculty of Psychology and Neuroscience of Maastricht University as a PhD student on verbal credibility assessment methods.

List of publications

INTERNATIONAL JOURNAL PUBLICATIONS

Bogaard, G., Meijer, E., Vrij, A., & Merckelbach, H. (2016). Strong but wrong: beliefs about verbal and nonverbal cues to deception. *PLoS ONE*, *1*, e0156615. doi:10.1371/journal.pone.0156615

Bogaard, G., Meijer, E. H., Vrij, A., & Merckelbach, H. (2016). Scientific Content Analysis (SCAN) cannot distinguish between truthful and fabricated accounts of a negative event. *Frontiers in Psychology, 7*, 243. doi: 10.3389/fpsyg.2016.00243

Bogaard, G., Meijer, E., & Vrij, A. (2014). Using an example statement increases information but does not increase accuracy of CBCA, RM, and SCAN. *Journal of Investigative Psychology and Offender Profiling, 11,* 151-163. doi: 10.1002/jip.1409

Bogaard, G., Meijer, E., Vrij, A., Broers, N. J., & Merckelbach, H. (2014). Contextual Bias in Verbal Credibility Assessment: Criteria Based Content Analysis (CBCA), Reality Monitoring (RM), and Scientific Content Analysis (SCAN). *Applied Cognitive Psychology, 28*, 79–90. doi: 10.1002/acp.2959

Bogaard, G., Meijer, E., Vrij, A., Broers, N. J., & Merckelbach, H. (2013). SCAN is largely driven by 12 criteria: Results from sexual abuse statements. *Psychology, Crime and Law*, *20*, 430-449. doi: 10.1080/1068316X.2013.793338

DUTCH JOURNAL PUBLICATIONS

Bogaard, G., Meijer, E., & Merckelbach, H. (2016). Screenen met SCAN? Liever niet. *Panopticon: Tijdschrift voor Strafrecht, Criminologie en Forensisch Welzijnswerk, 37*, 197-210.

Bogaard, G., Meijer, E., Vrij, A., & Merckelbach, H. (2011). Verbale analysemethoden: Leugenaars praten anders. *De Psycholoog*, 10-19.

DUTCH BOOKS

Bogaard, G., & Meijer, E. (2016). Verbale Leugendetectie wizards. *Politiekunde 80. Politie en Wetenschap*, Apeldoorn; Reed Business: Amsterdam, 1-88.

CONFERENCE PRESENTATIONS

Bogaard, G., et al. (June 2016). *Scientific Content Analysis (SCAN) cannot distinguish between truthful and fabricated accounts*. Paper presented at the European Association of Psychology and Law Conference (EAPL), Toulouse, France.

Bogaard, G., et al. (August 2015). *Strong but wrong: Beliefs about verbal and non-verbal cues to deception*. Paper presented at the Decepticon conference (International conference on deception research), Cambridge, United Kingdom.

Bogaard, G., et al. (June 2014). SCAN: *The popular lie detection method that cannot tell truth from lies*. Poster presented at the International Investigative Interviewing Research Group (IIIRG), Lausanne, Switzerland. **(Best poster award)**

Bogaard, G., et al. (June 2013). *Beliefs about verbal and non-verbal cues to deception*. Paper presented at the Forensic Psychology Update 2.0, Maastricht, The Netherlands.

Bogaard, G., et al. (May 2013). *Contextual bias in verbal credibility assessment methods*. Poster presented at the Association for Psychological Science (APS), Washington, United States.

Bogaard, G., et al. (April 2012). *SCAN criteria derived from field data*. Paper presented at the European Association of Psychology and Law Conference (EAPL), Nicosia, Greece.

Bogaard, G., et al. (July 2011). *Confirmation bias in verbal credibility assessment methods*. Paper presented at the Forensic Psychology Update, Maastricht, The Netherlands.

Bogaard, G., et al. (May 2011). *Confirmation bias in verbal credibility assessment methods.* paper presented at the Belgian Association for Psychological Sciences (BAPS), Gent.

Bogaard, G., et al. (June 2010). *Confirmation bias in verbal veracity assessment methods*. Poster presented at the European Association of Psychology and Law Conference (EAPL), Gothenburg, Sweden.