Contextual Bias in Verbal Credibility Assessment:

Criteria Based Content Analysis (CBCA), Reality Monitoring (RM), and Scientific Content Analysis (SCAN)

Glynis Bogaard[1], Ewout H. Meijer[1], Aldert Vrij[2], Nick J. Broers[1] and Harald Merckelbach[1]

[1]Maastricht University

[2]University of Portsmouth

*Requests for reprints should be addressed to Glynis Bogaard, Dept. of Clinical Psychological Science, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands (e-mail: Glynis.bogaard@maastrichtuniversity.nl).

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.

Keywords: lie detection, contextual bias, confirmation bias, Criteria Based Content Analysis, CBCA, Reality Monitoring, Scientific Content Analysis, SCAN

Contextual Bias in Credibility Assessment: Criteria Based Content Analysis (CBCA), Reality Monitoring (RM), and Scientific Content Analysis (SCAN)

People are not very successful in detecting lies. An extensive body of research shows that when they base their judgements on verbal and nonverbal behaviour, individuals, including trained police officers, generally perform only just above chance level (Aamodt & Custer, 2006; Bond & DePaulo, 2006, 2008; Elaad, 2003; Vrij, 2008). 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, 2008). Four studies examined SCAN, but found no solid evidence for its discriminative value (Driscoll, 1994; Nahari, Vrij, & Fisher, 2011; 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, Meijer, Vrij, Broers,

& Merckelbach, In press). 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 (Jones & Sugden, 2001; Findley & Scott, 2006). 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 or not 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 in 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?

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 real life files of the Amsterdam 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 A 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[1], 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

Since SCAN is an unstandardized method, we first investigated which SCAN criteria the SCAN trained police officers actually 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 B. 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 B for definitions). 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).

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 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%),“Social 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.

-Please insert Table 1 about here-

SCAN and contextual bias

To test whether the statements presented with positive context information scored higher in credibility compared to 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; d= 1.01). 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 to the control group, the use of SCAN did not mitigate the effects of extra-domain information on credibility ratings.

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 can not 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 et al., 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?

Experiment 1 suggested that SCAN may 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 additional 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 actually the third phase from 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, 2008). Although CBCA was developed for children, numerous studies have shown its usefulness with adult victim and/or eyewitnesses (Akehurst, Köhnken, & E, 2001; Sporer, 1997; Vrij, Akehurst, Soukara, & Bull, 2004; 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 taken into account, 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, 2008). 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, 2008). As to the individual criteria, the contextual (temporal

and spatial) criteria seem to have the highest diagnostic value (Masip, Sporer, Garido, & Herrero, 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. Unlike SCAN, as previously mentioned, SVA guidelines stress that the expert should gather as much information as possible about the case and about the person who wrote the statement (phase 1). Keeping in mind the order of the phases of SVA described above, a CBCA analyst has knowledge of all the contextual information gathered at phase 1 when analysing the statement. However, for the evaluation of the quality of a statement by use of CBCA, only the background of the victim's cognitive and verbal competence is necessary. The evaluation of other data (e.g., biographical information, behavioural information, etc.) is only necessary when making judgments about the complete overall credibility (SVA) (Steller, 1989). In sum, this means that 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, which 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 were 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.

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 specific, participants received information about the different criteria as described in chapters 8 to 10 by Vrij (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 example 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 C; 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 indicated deception (marked with * in C) . 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 in order 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 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.

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) = 150.4, *p* < 0.001; d = 1.1). 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) = .71, *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 3).

-Please insert Table 3 about here-

Criteria scores and contextual bias

To test whether participants actually 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; η_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 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 it was preceded by positive than when it was preceded by negative information, which reflects a profound contextual bias effect.

-Please insert Table 4 about here-

Individual criteria analyses

For the interested reader, Appendix D provides an overview of the Pearson correlations between the individual CBCA, RM and SCAN criteria, the total sum score and the associated credibility judgement. This shows to what extent the separate criteria contributed to the total sum score and to the credibility judgment.
Appendix E 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.

General Discussion

Are verbal credibility methods similarly sensitive to contextual bias? On the basis of 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) stated that the main purpose of credibility assessment is assessing the credibility of the statement and not the credibility of the witness. 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. 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.

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.

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

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

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

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.

Bogaard, G, Meijer, E, Vrij, A, Broers, Nick J, & Merckelbach, H. (In press). SCAN is largely driven by 12 criteria: Results from field data. *Psychology, Crime and Law*.

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.

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology, 74*, 271-280.

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.

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

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.

Elaad, E., Ginton, A., & Ben-Shakhar, G. (1994). The effects of prior expectations and outcome knowledge on polygraph examiners' decisions. *Journal of Behavioral Decision Making, 7*, 279-292.

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

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.

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

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

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. (2004). Statement validity ananlysis and the 'detection of the truth'. In P. A. Granhag & L. A. Strömwall (Eds.), *The detection of deception in forensic contexts*. Cambridge: University Press.

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.

Nahari, G., Vrij, A., & Fisher, R. P. (2011). Does the truth come out in the writing? SCAN as a lie detection tool. *Law and Human Behavior*, 1-11.

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

Porter, S., & Yuille, J. 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.

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.

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

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.

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.

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

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. (2008). *Detecting lies and deceit: Pitfalls and opportunities*. Chichester: Wiley.

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

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.

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

Williams, E. J. (1949). Experimental designs balanced for the estimation of residual effects of treatments. *Australian Journal of Scientific Research, 2*, 149-168.

Footnotes

[1] In a typical SCAN assessment, SCAN analysts start by asking the respondent to write down a 'pure version' of the event. This means that the respondent writes his/her own account of what happened without any interference from other people. Next, a copy of this pure version is matched against criteria mentioned in Appendix B. To indicate the presence of these criteria, SCAN experts primarily use colour codes. For example, marking a social introduction (my girlfriend, Amy) in green indicates its presence (Criterion 1, Appendix B).

Appendix A

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 behaviour 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 B

Since the SCAN manual does not consist of an overview of SCAN criteria, for purposes of this study we used an extensive list composed by four police officers from Amsterdam Amstelland who completed the SCAN course.

Description of SCAN criteria used in study 1.

1. *Social introduction* This criterion refers to how the persons in the statements are introduced. A proper introduction includes the name of the person and their role (e.g., "My husband, Eric…"). 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
2. *First person singular, past tense** This criterion refers to the format in which the statement is written. *:* 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. Deviations from past tense or writing in third person could indicate a lack of commitment, which, in turn, could indicate deception. For example, a statement written in first singular, present tense already fulfils the criterion, as one deviation is already present.
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.
4. *Use of pronouns** This criterion refers to the use of pronouns in the statement (e.g., "he", "my", "your" etc.). When the writer omits pronouns in the

statement this indicates an aversion of the writer to commit to the act described.

5. *Structure of the statement** This criterion refers to the balance of the statement. For each statement the number of lines is counted, next the lines of the statement are divided into a prologue, the main event and the epilogue. In a truthful statement it is expected that 20% of the lines are used to write the prologue (e.g., actions leading to the main event), 50% is used to write the main event, and 30% is used to write the epilogue (e.g., discussion about what happened after the event).
6. *Missing information** This criterion refers to the missing information in the statement. Missing information can be easily recognized when there are objective times in the statement. For example, "I arrived home at five o'clock, and started cooking at six o'clock". No information is given about what happened between five and six o'clock. Missing information indicates that the writer is (deliberately) hiding something.
7. *Out of sequence information** This criterion refers to the chronological order of the given information. When there is a deviation from the chronological order in the statement, this may indicate deception. This criterion also refers to information in the statement that does not seem relevant for the reader. This information is sensitive for the reader. In Vrij (2008) this criterion is taken together with the extraneous information criterion.
8. *Place of emotions* This criterion refers to the place where the emotions are present in the statement. SCAN suggests that emotions are located in unique places in the statement. It is expected that deceivers mention emotions before

the main event in the statement, whereas truth tellers are expected to mention emotions during or after the main event of the statement.

9. *Change in language** This criterion refers to a change of terminology or vocabulary in the statement. When there is a change in language, this means that something has changed in the mind of the author. It is possible that there is a justification for the change in language. In this case the story indicates truthfulness. If it is not possible to find a justification for the change in language, this change indicates deception.
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.
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.
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.
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.
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.
15. *Objective versus subjective time* This criterion refers to the relationship between subjective and objective time. Subjective time refers to the amount of

text written by the author to describe an event, whereas objective time refers to the actual time the event prolonged. In a truthful statement it is expected that the subjective time corresponds to the objective time. For example, if a writer uses 2 lines to describe 15 minutes, then he/she should use 4 lines to describe 30 minutes.

16. *Extraneous information** This criterion refers to information that does not seem relevant for the reader. It is expected that a writer includes extraneous information to hide something else. Therefore, extraneous information may indicate deception.
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.
18. *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.
19. *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 (Police AMS).
20. *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 (Police AMS).

21. *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. (Police AMS). In Vrij (2008) this criterion is a combination of "Denial of allegation" and "Lack of conviction or memory"

Appendix C

Description of criteria used in study 2

CBCA criteria (Steller & Köhnken, 1989)

1. Logical structure
2. Unstructured production
3. Quantity of details
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
13. Attribution of perpetrator's mental state
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

RM criteria (Sporer, 1997)

1. Clarity

2. Perceptual information
3. Spatial information
4. Temporal information
5. Affect
6. Reconstructability of the story
7. Realism
8. Cognitive operations

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 D

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 E

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		*t*	Effect size (r)
		M	SD	M	SD		
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 C. *). t refers to the t-value of the difference between scores in the Positive and Negative information condition.*

Table 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

Table 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.

Table 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.

Condition	Negative		Positive	
	M (*SD*)	Skewness (*SE*)	M (*SD*)	Skewness (*SE*)
Control	2.89 *(.83)*	0.37 *(.41)*	4.84 *(.0.94)*	-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 *(.0.98)*	0.06 *(.41)*	4.36 *(1.13)*	-.73 *(.41)*

Table 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.

Method	Negative		Positive	
	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)*